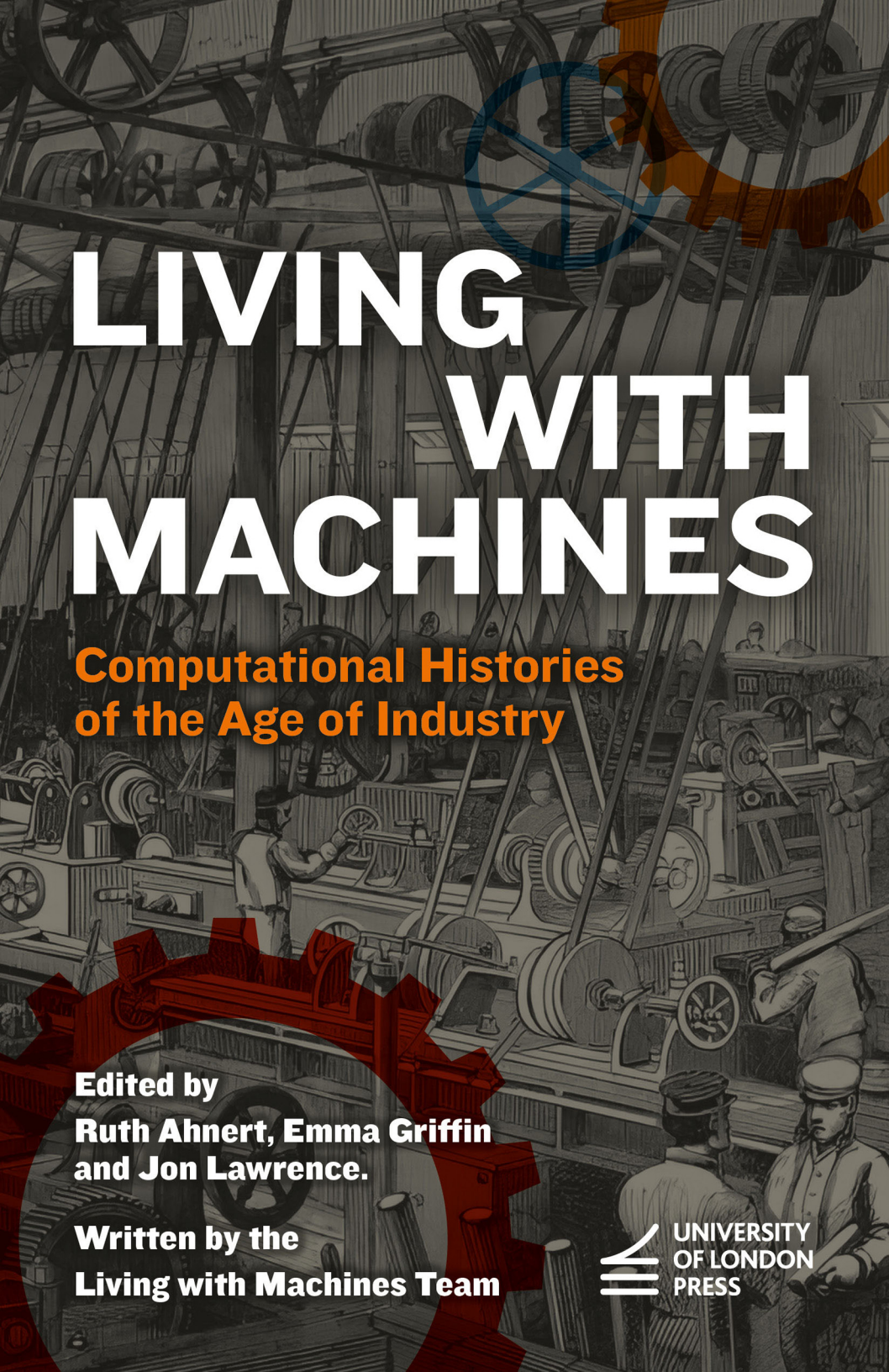


LIVING WITH MACHINES

Computational Histories of the Age of Industry

Edited by
Ruth Ahnert, Emma Griffin and Jon Lawrence.

Written by the
Living with Machines Team

UNIVERSITY OF LONDON PRESS

Living with Machines

Living with Machines

Computational Histories of the
Age of Industry

Edited by

Ruth Ahnert, Emma Griffin and Jon Lawrence

Written by the Living with Machines team
Ruth Ahnert, Kaspar Beelen, Arianna Ciula,
Emma Griffin, Timothy Hobson, Kasra Hosseini,
Katherine McDonough, Barbara McGillivray,
Jon Lawrence, Sherman Lo, Tiffany Ong,
Nilo Pedrazzini, Joshua Rhodes, Mia Ridge,
Guy Solomon, Daniel van Strien, Miguel Vieira,
Valeria Vitale, Kalle Westerling,
Daniel C.S. Wilson and Rosie Wood

UNIVERSITY
OF LONDON
PRESS

Available to purchase in print or download
for free at https://uolpress.co.uk

First published 2026 by
University of London Press
Senate House, Malet St, London WC1E 7HU

© The Authors 2026

The right of the authors to be identified as authors of this Work has been asserted by them in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

This book is published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.



Any third-party material reproduced in the book is not covered by the book's Creative Commons licence, unless indicated otherwise in the image or text's credit line. To reuse third-party material not published under the same licence as the book you will need to obtain permission from the copyright holder.

A CIP catalogue record for this book is available from The British Library.

ISBN 978-1-914477-63-8 (hardback)
ISBN 978-1-914477-64-5 (paperback)
ISBN 978-1-914477-67-6 (.epub)
ISBN 978-1-914477-66-9 (.pdf)
ISBN 978-1-914477-65-2 (.html)

DOI https://doi.org/10.14296/xdyp6338

Cover image: The Lathe and Tool Shop, Stephenson's Locomotive Manufactory at Newcastle-on-Tyne, *The Illustrated London News*, vol. 45 (15 October 1864), 393.

Cover design for University of London Press by Nicky Borowiec.
Book design by Nigel French.
Text set by Westchester Publishing Services UK in Meta Serif and Meta, designed by Erik Spiekermann.

Contents

List of figures vii

List of tables xv

Authorship, credit and acknowledgements xvii

Introduction: working with machines 1
Ruth Ahnert, Emma Griffin and Jon Lawrence

1. **The Environmental Scan as a method for digital source criticism at scale** 33
Kaspar Beelen, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Daniel van Strien and Daniel C.S. Wilson

2. **MapReader: software and principles for computational map studies** 91
Katherine McDonough, with Ruth Ahnert, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Valeria Vitale, Kalle Westerling, Daniel C.S. Wilson and Rosie Wood

3. **Beyond the tracks: re-connecting people, places and stations in the history of late Victorian railways** 131
Joshua Rhodes, Jon Lawrence, Kaspar Beelen, Katherine McDonough and Daniel C.S. Wilson

4. **Links in the chain: workers in the British bicycle boom, as seen from the England and Wales census, 1881–1911** 179
Guy Solomon, Joshua Rhodes, Jon Lawrence, Emma Griffin, Timothy Hobson, Sherman Lo and Daniel C.S. Wilson

5. **Analysing the language of mechanisation in nineteenth-century British newspapers** 255
Barbara McGillivray, Nilo Pedrazzini, Arianna Ciula, Jon Lawrence, Tiffany Ong, Mia Ridge and Miguel Vieira

Afterword: historians and the future of AI 305
Ruth Ahnert

Bibliography 323

Index 357

List of figures

Figure 1.1a: Front page of the Mitchell's *Newspaper Press Directory* from 1908 42

Figure 1.1b: An example of a newspaper description of the *Chatham News* in Chatham (Kent) 43

Figure 1.2: Issues with historical fonts and Optical Character Recognition 44

Figure 1.3: Image-processing pipeline 45–46

Figure 1.4: Distribution of political labels 51

Figure 1.5: Topics positioned by the strength of their association with liberal (left) or conservative (right) 53

***Figure 1.6:** Ordnance Survey of Great Britain, William Mudge, Thomas Colby, Benjamin Baker, *Index to the Ordnance Survey of England and Wales, Published in Sheets on a Scale of one inch to a Mile* (Ordnance Survey, 1870) 62

Figure 1.7: National Library of Scotland's 'Map Finder' interface 63

***Figure 1.8:** Louis Capitaine, César-François Cassini, and the Cassini family, *Tableau de la carte générale de la France* (L. Capitaine, 1797) in César-François Cassini and the Cassini family, *Carte de France* (L. Capitaine, 1750–1815) 64

Figure 1.9: The National Library of Scotland 'Map Records Search' interface showing 25-inch sheet bounding boxes, with specific sheets printed over time and available in the NLS collections displayed in the top right-hand corner 66

***Figure 1.10:** Detail from Macromap of three scales of OS maps plotted over time based on their earliest available survey or print date 68

Figure 1.11: Detail from Macromap showing sheet link in the bottom of the Control Panel, which appears after selecting the sheet's location on one of the mini maps to the left 69

***Figure 1.12:** Olivia Vane, *OS 'Old Series'/'First Series', England & Wales* (2023) 70

***Figure 1.13:** Olivia Vane, *OS 6"-to-the-mile, Great Britain: 2nd Ed.* (2021) 71

Figure 1.14a: Relative frequency of the words 'travel(s)/travelling' in book titles 75

Figure 1.14b: Relative frequency for the words 'machine(s)' and 'engine(s)' in book titles 76

Figure 1.15: The proportion of metadata annotated ('true') versus not annotated ('false') in the BL Microsoft Books collection 77

Figure 1.16: The proportion of fiction works in the Microsoft Books Corpus (orange) and British Library Catalogue (blue) 79

Figure 2.1: Examples of MapReader patches (with unique ID labelled above) from historical Ordnance Survey mapping 97

***Figure 2.2:** Ordnance Survey, Six-inch to one-mile maps of England, Wales and Scotland, 2nd Edition – 1887–1949. Animated chronology by publication year 102

Figure 2.3: Seven MapReader steps for patch-classification pipeline 103

Figure 2.4: Screenshot of annotation step with context turned off 104

Figure 2.5: Patch centroids for railspace label in south Lancashire, near Wigan 106

Figure 2.6: MapReader's text-spotting pipeline 108

Figure 3.1: Map displaying both population density and railway lines, used by Geddes, (original from Bartholomew's *Royal Atlas of England*, 1899) 134

Figure 3.2: 'Over-London by rail', Gustav Doré print, from *London: A Pilgrimage* (Blanchard Jerrold, London, 1872) 137

Figure 3.3: Detail of MapReader results for railspace patches (black) around Leeds 146

Figure 3.4: Percentage of households with live-in servants by distance to station and parish type, Great Britain (1901) 149

Figure 3.5: Percentage of urban households with live-in servants by distance to station and railspace density, Great Britain (1901) 151

Figure 3.6: Percentage of rural households with live-in servants by distance to station and railspace density, Great Britain (1901) 151

Figure 3.7: Percentage of all urban occupied workers in custom occupational groups by distance from nearest station, Great Britain (1901) 153

Figure 3.8: Percentage of workforce in custom occupational groups by railspace density, Great Britain (any distance from station), 1901 154

Figure 3.9: Percentage of workforce in custom occupational groups by railspace density (urban only; living less than 2 km from station), Great Britain, 1901 155

Figure 3.10: Mean railspace and distance to station by parish, Great Britain (1901) 158

Figure 3.11: Map of spatial distribution of servant-keeping in relation to railspace and stations, Upper Holloway (London), 1901 159

Figure 3.12: Map of spatial distribution of central business district workers in relation to railspace and stations, Upper Holloway (London), 1901 160

Figure 3.13: Map of spatial distribution of servant-keeping in relation to railspace and stations, Camberwell (London), 1901 162

Figure 3.14: Map of spatial distribution of central business district workers in relation to railspace and stations, Camberwell (London), 1901 162

Figure 3.15: Map of spatial distribution of servant-keeping in relation to railspace and stations, Middlesbrough, 1901 164

Figure 3.16: Map of spatial distribution of central business district workers in relation to railspace and stations, Middlesbrough, 1901 164

Figure 3.17: Map of spatial distribution of servant-keeping in relation to railspace and stations, Salford (Manchester), 1901 166

Figure 3.18: Map of spatial distribution of central business district workers in relation to railspace and stations, Salford (Manchester), 1901 166

***Figure 4.1:** The distribution of bicycle makers, by census year: a) 1881, b) 1891, c) 1901, d) 1911 186–189

***Figure 4.2:** The distribution of composite metalworkers, by census year: a) 1881, b) 1891, c) 1901, d) 1911 190–193

***Figure 4.3:** Linked bicycle makers relative to all workers in this occupation, 1881 to 1911 195

***Figure 4.4:** Linked composite metalworkers relative to all workers in this grouping, 1881 to 1911 196

***Figure 4.5:** Bicycle makers linked to the previous census: a) 1891 to 1881, b) 1901 to 1891, c) 1911 to 1901 199–201

***Figure 4.6:** Composite metalworkers linked to the previous census: a) 1891 to 1881, b) 1901 to 1891, c) 1911 to 1901 204–206

***Figure 4.7:** Bicycle makers linked to the previous census: a) 1891 to 1881, Coventry; b) 1891 to 1881, Birmingham; c) 1901 to 1891, Coventry; d) 1901 to 1891, Birmingham; e) 1911 to 1901, Coventry; f) 1911 to 1901, Birmingham 207–212

***Figure 4.8:** Composite metalworkers linked to the previous census: a) 1891 to 1881, Coventry; b) 1891 to 1881, Birmingham; c) 1901 to 1891, Coventry; d) 1901 to 1891, Birmingham; e) 1911 to 1901, Coventry; f) 1911 to 1901, Birmingham 214–219

***Figure 4.9:** The occupation of household heads, where a member of the household had no occupation and was a bicycle maker in the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911 221–223

***Figure 4.10:** The occupation of household heads, where a member of the household had no occupation and was a bicycle maker living in Coventry at the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911 224–226

***Figure 4.11:** The occupation of household heads, where a member of the household had no occupation and was a composite metalworker in the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911 227–229

***Figure 4.12:** The occupation of household heads, where a member of the household had no occupation and was a composite metalworker living in Coventry at the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911 230–232

***Figure 4.13:** Bicycle makers linked to the following census: a) 1891 to 1901, b) 1901 to 1911 233–234

***Figure 4.14:** Composite metalworkers linked to the following census: a) 1891 to 1901, b) 1901 to 1911 236–237

***Figure 4.15:** Bicycle makers in 1901 linked to the 1911 census in a) Coventry, b) Birmingham 238–239

Figure 5.1: Flowchart of the methods for the automatic detection of meaning change with diachronic word embeddings, from pre-processing to time-series analysis 260

Figure 5.2: Example of word embeddings as points in space, produced from the word embedding model trained on the 1880s newspaper subcorpus 262–263

Figure 5.3: Examples of vector representations of four words from the diachronic word embedding models trained on the 1860s and 1870s newspaper subcorpora 264

Figure 5.4: The ten nearest neighbours to 'machine' in the 1860s pre- (left) and post- (right) OCR-error correction 265

Figure 5.5: Automatically detected changepoints in the semantic change trajectory of a) 'train' and b) 'railway' 266

Figure 5.6: Comparison between the semantic change trajectories of three words: 'gear', 'railway' and 'train' 267

Figure 5.7: Application of a time-series clustering technique (k-means with Dynamic Time Warping) to semantic change trajectories 268

Figure 5.8: Semantic change trajectory of 'train' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century 269

Figure 5.9: Semantic change trajectory of 'railway' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century 270

Figure 5.10: Semantic change trajectory of 'gear' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century 271

Figure 5.11: Semantic change trajectory of 'coach' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century 271

Figure 5.12: Semantic trajectory of 'cars' in the North of England, based on the nearest neighbours of this word in the respective geographic subcorpus in subsequent decades 273

Figure 5.13: Semantic trajectory of 'cars' in the South of England, based on the nearest neighbours of this word in the respective geographic subcorpus in subsequent decades 274

Figure 5.14: The Zooniverse task interface showing newspaper articles and the related annotation options for 'coach' 276

Figure 5.15: Raw total count of annotations for each word meaning ('trolley' as cart vs. trolley-car) in each year 277

Figure 5.16: Raw total count of annotations for each word meaning ('trolley' as cart vs. trolley-car vs. other) and other classifications in each year 278

Figure 5.17: Relative frequencies of the annotations for each of the word meanings for the word 'trolley' 279

Figure 5.18: Decades when the words a) 'trolley', b) 'bike' and c) 'coach' may have changed meaning according to our changepoint detection analysis 280

Figure 5.19: Raw total count of annotations for each word meaning associated to a) 'bike' (powered by human legs vs. an engine), b) 'coach' (horse-drawn carriage vs. motor bus vs. railway carriage) and c) 'car' (railway vs. road vehicle powered by a motor vs. wheeled horse-drawn conveyance vs. electric tramcar) respectively, in each year 283

Figure 5.20a: Relative frequencies of the annotations for each of the word meanings for 'bike' in each year 284

Figure 5.20b: Relative frequencies of the annotations for each of the word meanings for 'coach' in each year 284

Figure 5.21: Decades when the words a) 'bike', b) 'coach' and c) 'car' respectively may have changed meaning according to our changepoint detection analysis (window set at 2 and penalty at 25) 285

Figure 5.22: Relative frequencies of the annotations for each of the word meanings for 'car' in each year 286

Figure 5.23: Potential decades when the word 'car' changed meaning with two change points represented by the vertical lines (penalty decreased to 22) 286

Figure 5.24: Number of articles per year, from 1800 to 1920 291

Figure 5.25: Political leaning of the newspaper titles from which we sourced the articles 291

Figure 5.26: Number of annotators that contributed to each workflow 291

Figure 5.27: Level of agreement around the meaning of the word 'trolley' across the volunteer annotators 292

Figure 5.28: Level of inter-annotator agreement around the meaning of the words a) 'car', b) 'bike' and c) 'coach' respectively 292

Figure 5.29: Example of a 'subject' – an image taken from an 1864 article where the word 'trolley' is highlighted 294

Figure 5.30: Another subject, an image of part of an article from 1903 where the word 'trolley' is highlighted 294

Figure 5.31: Simplified data model of Zooniverse data 295

Figure 5.32: Screenshot of static mock-ups sketched in Miro board during the Discovery phase 296

***Interactive versions of these figures are included in the open access edition of the book on the University of London Press instance of the Manifold platform:** https://read.uolpress.co.uk/projects/living-with-machines. This edition also includes supplementary resource collections from the project including publications; datasets and new digitised content; software, tools, pipelines, and models; videos from the Living with Machines documentary series; and data visualisations.

List of tables

Table 1.1: Selected examples from the structured *Newspaper Press Directories* showing the extracted data for the *Brackley Observer* and *Altrincham and Bowdon Guardian* 50

Table 1.2: Frequency of the ten most common shifts in political orientation 52

Table 1.3: Evaluation results for the fiction/non-fiction classifier 78

Table 3.1: Sample I-CeM data 141

Table 3.2: Number and percentage of individuals geo-coded in 1901 Great Britain Census 143

Table 3.3: Number and percentage of individuals by railspace density, 1901 Great Britain (within 100 m from street) 150

Table 3.4: Number and percentage of individuals by urban/rural classification, Great Britain, 1901 169

Table 3.5: Number and percentage of individuals at different distances from stations by urban class 170

Table 3.6: Custom I-CeM occupational coding 171

Table 4.1: Nominal linkage scenarios 194

Table 4.2: Number of bicycle workers, 1881–1911 244

Table 4.3: Occupations in our 'composite metalworker' classification 246

Authorship, credit and acknowledgements

This is not an edited collection but a co-authored book by the Living with Machines team. All members of the team who contributed to the writing of this book are listed in full on the title page. The cover bears the additional names of Ruth Ahnert, Emma Griffin and Jon Lawrence to acknowledge their editorial labour, shaping and overseeing the delivery of the book as a whole. However, the book is a larger collective effort, which emerges from a five-year project, Living with Machines (2018–23), which was funded by UK Research and Innovation's Strategic Priorities Fund and delivered by the Arts and Humanities Research Council (AHRC). The book therefore sits atop a huge quantity of shared thought and labour that is hard to acknowledge in a standard model of authorship. The project team collaboratively and iteratively co-created every dimension of this project: its values of collaboration, the ways of working, the methods, and the intellectual questions and ambitions on which we set our sights. As such, the book has been shaped by all forty-six members, and the numerous other formal and informal collaborators that passed through during the project's five-year duration, even if they do not appear as authors. For this reason we list all members of the project below.

A little more should be said, however, about how we have chosen to credit labour in the individual chapters. Crediting a collective endeavour is complex. We might look to the Multigraph Collective as a comparison. The Multigraph Collective is a team of twenty-two scholars at sixteen universities in Canada, the US and the UK. In the book that they co-authored, the use of their collective name acts to share credit equally among its members.[1] But a collective name can also obscure the identities of its members (sometimes intentionally, as with the Xenofeminist collective Laboria Cuboniks, before the members ultimately identified themselves)[2] and also the size and nature of contributions. In some intellectual collaborations this might not matter: in the case of the Multigraph Collective, they describe their writing process, how each person contributed text (roughly amounting to a third of a chapter) and how these were then 'winnowed' and 'grafted' together into a coherent and multifaceted text.[3] However, what's

important to note is that their contributions were all of one kind: thinking and writing.

The difference with our project is that the contributions are of very different kinds, because of the highly interdisciplinary nature of the work. These vary from ensuring we have the rights to access particular datasets, to the ingest and wrangling of data, the design of databases and data structure, the development of new analytical methods, the parsing of results, the contextualising of those results within historical context, and the construction of all this work into the texts of the chapters. We believe that all these types of labour deserve credit, and that the best way to do this is to be transparent about whose labour went where. This is also important because there are radically different conventions for crediting intellectual labour across the multiple disciplinary and professional cultures that the Living with Machines team represents. Moreover, this kind of accountability is important to those collaborators who may need to evidence their contributions, such as, for example, people earlier in their careers who are building their CVs.

This is the reason that our author lists on each chapter are longer than is the norm in history publications, or indeed those in many other humanities disciplines. Our general policy as a team is that all parts of the research workflow should be credited through either citation or an author credit.[4] In practice this means that if there is an earlier publication by some members of the team that reports on a prior part of that workflow, this can be cited, and the result may be a shorter author list for the current publication. However, if there is no prior publication, all those who contributed (at whatever stage) will be named as authors.

We also seek to describe accurately the different types of contribution. Following the common practice in the sciences, we denote lead authorship by placing that author in first (or joint-first) place in the author-list for each chapter. However, we also wanted to make clear what kind of work that leadership denotes, as well as the size and kind of contribution of the other authors. As such we accompany our chapters with a statement of contribution – now a common practice in many science journals – which appears in the first footnote of each. These statements are loosely based on CRediT (Contributor Roles Taxonomy), which is a high-level taxonomy comprising fourteen roles that can be used to represent the labour typically contributed by contributors to research outputs.[5] These needed to be adapted to the particulars of historical analyses using cultural heritage

collections, so the statements found in the pages following seek to identify activities such as conceptualisation, data curation, method design, implementation, reproducibility, interpretation, historical analysis, annotation, writing and editing, supervision and project management.

We also made the decision that principal investigator and co-investigators would not automatically be authors on all publications. This was the right decision for us in this particular context, but it is not necessarily one that works for other fields or even other digital humanities projects. Moreover, it is a decision that shows the difficulty of drawing lines around what labour does or does not count. It does, however, follow the lead being taken by various scientific journals, which are requiring that everyone listed as an author should meet their criteria for authorship. The framing of such policies seems to be designed to ward against principal investigators or lab leads automatically gaining an author credit, regardless of whether they made a substantive intellectual contribution or not. This is a practice that has become fairly standard in many scientific spaces. The International Committee of Medical Journal Editors' (ICMJE) Uniform Requirements for Manuscripts Submitted to Biomedical Journals is a good example of a policy that seeks to insist on a substantive contribution. It lists four conditions for authorship credit that all authors must meet:

- Substantial contributions to conception and design; or the acquisition, analysis, or interpretation of the data for the work: AND
- Drafting the work or reviewing it critically for important intellectual content; AND
- Final approval of the version to be published: AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.[6]

We have followed a similar spirit of authorship on the project more generally as it seemed like the fairest and most transparent policy overall. We might think of it as inclusivity without bloat.

The one downside of where we chose to draw the line is that it obscures certain forms of labour that are crucial for creating the environment necessary for research and publication, even if they do not directly contribute to a specific output. In particular, there was a whole layer of labour that was required for creating the research environment in which this book could be

written, and the research on the larger project could take place. This begins with grant capture but continues with financial management, project management, supervision, strategising, internal and external reporting, steering the research vision, overseeing legal data access and legal agreements. This labour was provided by our project manager André Piza (and his parental leave cover Karen Cordier), project coordinator Léllé Demertzi (who also documented and formatted the supplementary resources for this book), principal investigator Ruth Ahnert (and her maternity leave cover Emma Griffin and David Beavan), rights assurance manager Claire Austin, and Mia Ridge and Maja Maricevic at the British Library – only some of whom bear author credits on this book. James Hetherington, Adam Farquhar and Giovanni Colvizza shaped the project in crucial ways at inception, and in their roles as co-investigators before their departures. James, in particular, introduced the wider team to software engineering research practices that materially shaped the way we worked as a team, as well as making sure that the research software engineers and data scientists were equal collaborators and not merely service providers, as they are treated on so many projects. There are various other pieces of work provided by the research software engineers (RSEs) on the project that aren't credited here with authorship because of their foundational function (such as early newspaper data wrangling by Sarah Gibson), because the work led to other publications and outputs not featured here (such as the work on numerous Natural Language Processing pipelines and outcomes by Fede Nanni, or the work on the newspaper metadata database by Griff Rees), or because the work came earlier in the pipeline and is cited in the chapters.

In other words, in a different model of credit, our author lists could have been longer still. In recognition of this, here follows a full list of all members of the project, and formal collaborators, during the period from 1 September 2018 to 31 July 2023 (later entries and departures are signalled in brackets).

- Ruth Ahnert, Principal Investigator, Queen Mary University of London; now University of Cambridge.
- David Beavan, Co-Investigator (research software engineer from September 2018, Co-Investigator from June 2019, acting joint-Principal Investigator December 2021–June 2022), Alan Turing Institute.

- Giovanni Colavizza, Co-Investigator, Alan Turing Institute (September 2018–June 2019); now University of Copenhagen.
- Adam Farquhar, Co-Investigator, British Library (September 2018–September 2019).
- Emma Griffin, Co-Investigator (acting joint- Principal Investigator December 2021–June 2022), Queen Mary University of London (QMUL).
- James Hetherington, Co-Investigator, Alan Turing Institute (September 2018–January 2020); now University College London (UCL).
- Timothy Hobson, Co-Investigator (research software engineer from November 2018, Co-Investigator from February 2020 until August 2023), Alan Turing Institute.
- Jon Lawrence, Co-Investigator, University of Exeter.
- Maja Maricevic, Co-Investigator (Co-Investigator from September 2019), British Library.
- Barbara McGillivray, Co-Investigator, King's College London.
- Mia Ridge, Co-Investigator, British Library.
- Sir Alan Wilson, Co-Investigator, Alan Turing Institute.
- Claire Austin, Rights Assurance Manager, British Library (November 2018–July 2023).
- Kaspar Beelen, Digital Humanities Senior Research Associate, Alan Turing Institute (February 2019–April 2023); now Institute of Advanced Studies, University of London.
- Felipe Bento, British Library Labs Technical Lead, British Library (November 2021 to April 2022).
- Mariona Coll-Ardanuy, Computational Linguistics Senior Research Associate, Alan Turing Institute (January 2019–September 2022 and March 2023–July 2023); now Barcelona Supercomputing Center.
- Arianna Ciula, Director of King's Digital Lab and Senior Research Software Analyst (external collaborator, February 2023–July 2023), King's College London.
- Karen Cordier (acting Project Manager December 2020–March 2021), Alan Turing Institute.
- Joel Dearden, Research Software Engineer, Alan Turing Institute (February 2019–July 2020 – *in memoriam*).
- Léllé Demertzi, Project Administrator (June 2022–July 2023), Alan Turing Institute.

- Rosa Filgueira, Data Architect, University of Edinburgh (external collaborator January 2019–September 2019); now University of St Andrews.
- Sarah Gibson, Research Software Engineer, Alan Turing Institute (May 2020–May 2021); now 2i2c.
- Luke Hare, Research Data Scientist (July 2022–July 2023), Alan Turing Institute.
- Lucy Havens (Visiting Researcher at Alan Turing Institute, June–September 2022), University of Edinburgh.
- Kasra Hosseini, Research Data Scientist, Alan Turing Institute (April 2019–January 2022); now Zalando Technology.
- Michael Jackson, Software Architect (external collaborator January 2019–July 2019), University of Edinburgh.
- Amy Krause, Data Architect (external collaborator June 2019–March 2020), University of Edinburgh.
- Christina Last, Research Data Scientist, Alan Turing Institute (July 2021–May 2022); now Massachusetts Institute of Technology.
- Sherman Lo, Research Software Engineer (August 2022–July 2023), Queen Mary University of London.
- Katherine McDonough, History Senior Research Associate, Alan Turing Institute (March 2019–July 2023); now Carnegie Mellon University.
- Federico Nanni, Research Data Scientist (November 2019–May 2023), Alan Turing Institute.
- Tiffany Ong, Research Software Engineer (external collaborator, February–July 2023), King's College London.
- Nilo Pedrazzini, Research Associate in Corpus-Based Digital Humanities, Alan Turing Institute (January 2022–July 2023), now QMUL.
- André Piza, Research Project Manager, Alan Turing Institute; now Goldsmiths University.
- Griffith Rees, Research Data Scientist (June 2022–July 2023), Alan Turing Institute.
- Joshua Rhodes, Senior Research Associate, Alan Turing Institute (November 2020–December 2022); now UCL.
- Yann Ryan, Curator of Digital Newspapers, British Library (external collaborator, January 2019–January 2020), now Leiden University.

- Andrew Smith, Research Data Scientist (January 2022–March 2023), Alan Turing Institute.
- Guy Solomon, Research Associate, Alan Turing Institute (April 2022–December 2022); now University of Sheffield.
- Giorgia Tolfo, Data and Content Manager, British Library (January 2019–February 2023); now The National Archives.
- Daniel van Strien, Digital Curator, British Library (February 2019–December 2022); now Hugging Face.
- Olivia Vane, Digital Humanities Research Software Engineer, British Library (June 2019–November 2021); now *The Economist*.
- Miguel Vieira, Research Software Engineering (external collaborator January–July 2023), King's College London.
- Kalle Westerling, Research Software Engineer, British Library (January 2022–June 2023); now Chainguard.
- Daniel C.S. Wilson, History Senior Research Associate, Alan Turing Institute (December 2018–July 2023); now UCL.
- Rosie Wood, Junior Research Data Scientist (January 2023–July 2023), Alan Turing Institute.

Beyond the formal collaborators we have an extensive network of people we would like to acknowledge. Our advisory board provided invaluable guidance and expertise at key stages in the project: Chair, Martin Daunton (University of Cambridge), and members Melodee Wood (Loughborough University), Adam Farquhar (independent, formerly British Library), Tessa Hauswedell (UCL), Dr James Hetherington (UCL), Edward Higgs (University of Essex), Tim Hitchcock (University of Sussex), Mary McKee (FindMyPast), Paul Meller (*ex officio*, AHRC), Aoife O'Connor (FindMyPast), Andrew Prescott (University of Glasgow), Tom Rodden (*ex officio*, Department for Culture, Media and Sport), David de Roure (University of Oxford), Jon Rowe (Alan Turing Institute/University of Birmingham), James Smithies (University of Canterbury, Australia), Alan Sudlow (*ex officio*, AHRC), Melissa Terras (University of Edinburgh) and Jane Winters (School of Advanced Study, University of London). The AHRC played a more active role in the oversight of this project than is standard on smaller projects that they fund. For their support and input we would especially like to thank Tao-Tao Chang, Dawn Greenberg, Deborah Kroll and current and former executive chairs Christopher Smith and Andrew Thompson.

We would not have been able to undertake this project without the institutional support of the British Library and the expertise and labour of numerous staff members past and present, especially in the imaging studio, Maps and News curators, the Digital Scholarship and Heritage Made Digital teams, and BL Labs. Special mentions to Beth Gaskell, Tom Harper, Luke McKernan, Victoria Morris, Gethin Rees, Saba Syed, Karl Harris, Graham Jevon, Neil Fitzgerald, Andy Appleyard, Matt Lambert, and the former chief executive Roly Keating. We are indebted to John McGoldrick and Rachael Dilley at Leeds Museums and Galleries for their generous collaboration on our Living with Machines exhibition at Leeds City Museum. Similarly, we are grateful for the exciting interdisciplinary environment provided by our home in the Alan Turing Institute, and specifically for the practical support and astute insights of Jonathan Atkins, Darren Grey, Nico Guernion, Catherine Lawrence, Amit Mulji, Martin O'Reilly, Jon Rowe, Kirstie Whittaker, Hua Yang and former chief executive officers Adrian Smith and Sir Alan Wilson. We are indebted to the wider collaborative team at King's Digital Lab who supported work on Chapter 5, beyond the listed authors, including Paul Caton, Shalen Fu, Elliott Hall, Pam Mellen, and Geoffroy Noël. Map images and data attributed to the National Library of Scotland (NLS) have been sourced using the NLS's Map Images website: https://maps.nls.uk/, and we express our special thanks to map curator Chris Fleet for his support both practical and intellectual.

We are eternally grateful to Emma Gallon and Jamie Bowman at University of London Press, for working with us to turn our vision for this book into a reality. We are in the unique position of being able to thank our reviewers – Alison Booth, Ryan Cordell, Paul Fyfe, Ian Milligan and Andrew Prescott – by name, thanks to the open nature of the process. The book is all the better for their astute comments; any remaining faults are, of course, our own.

Notes

1 The Multigraph Collective, *Interacting with Print: Elements of Reading in the Era of Print Saturation* (University of Chicago Press, 2018). The Multigraph Collective's members are Mark Algee-Hewitt, Angela Borchert, David Brewer, Thora Brylowe, Julia Carlson, Brian Cowan, Susan Dalton, Marie-Claude Felton, Michael Gamer, Paul Keen, Michelle Levy, Michael Macovski, Nicholas Mason, Nikola von Merveldt, Tom Mole, Andrew Piper, Dahlia Porter, Jonathan Sachs, Diana Solomon, Andrew Stauffer, Richard Taws and Chad Wellmon.

2 https://laboriacuboniks.net/.

3 See *Interacting with Print*, preface.

4 For further discussion, see Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), chapter 4.3. https://doi.org/10.1017/9781009175548.

5 See https://credit.niso.org/.

6 www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Introduction: working with machines

Ruth Ahnert, Emma Griffin and Jon Lawrence[1]

In 2023 – the final year of our five-year, multidisciplinary project 'Living with Machines' – Open AI launched ChatGPT. AI had arrived in the consciousness of the general public. Within months, the world had become gripped by the question of whether machines were about to make us all redundant, and, in more doom-laden scenarios, whether humanity itself was about to be overthrown by a new generation of super-intelligent machines.[2] There is nothing new about a technological innovation generating wild speculation about the future of society, including radical predictions of a future where no one would *need* to work (in the optimists' telling), or (in more apocalyptic versions), a future where most people would cease to have either purpose or any reliable source of remuneration. Similar things had been said in the nineteenth century about the rise of machines to replace hand labour, and in the twentieth century about first mass production methods, and, later in the century, about the rollout of robots and other automated manufacturing processes. In the 1970s and 1980s, social scientists debated the coming post-employment society that would be ushered in by automation and the off-shoring of industry.[3] They were right to anticipate the devastating social consequences that would be wrought by deindustrialisation, especially in political systems wedded to the idea that markets would provide their own solutions, but worklessness did not become the new norm.[4] We cannot know what changes the AI revolution will bring. They may yet prove to be more radical than any preceding technological transformations, but it is important to remember that much of the hyperbole about AI has been driven by people within the industry who

have every reason, both professionally and commercially, to play up the transformative power of their products.[5]

The impact of these technological changes on society have been described as the 'fourth industrial revolution',[6] a label that makes clear the connection of our current economic moment to the coming of the machine age in the nineteenth century, or 'industrial revolution' as it came to be called.[7] There is, in addition, a continuity in the cultural anxiety that accompanies technological change. If we turn from today's anxiety about AI bots to nineteenth-century commentary about the steam engine, we encounter a familiar set of fears about the dangers of new technology. William Blake's evocative lines about the 'dark Satanic Mills'; Thomas Carlyle's warnings about the arrival of the 'Mechanical Age' – where men have grown 'mechanical in head and in heart, as well as in hand';[8] William Wordsworth's critique of the 'outrage done to nature' by the growth of urban industry;[9] or the industrial novels of Benjamin Disraeli, Elizabeth Gaskell, Charles Dickens and, from later in the century, George Gissing,[10] all provide echoes of the concern that technological progress carried a high cost. The Industrial Revolution did not merely 'revolutionise' industry; it brought change to the landscape and environment, to social and gendered norms, to attitudes to money, to labour and the practice of everyday life. Furthermore, the specific anxiety was often expressed that mechanisation – through some combination of de-skilling, downward pressure on wages and unemployment – undermined the position of those who work for a living. As John Stuart Mill wrote, society's mechanical inventions had done no more than 'enable a greater proportion to live the same life of drudgery and imprisonment'.[11] For Victorian commentators, the conclusion was that technological progress upset the social order, not only for those at the bottom of the social order but also in the way that 'new money' threatened to displace landed wealth.

It was not simply nineteenth-century contemporaries who voiced concern about the impact of mechanisation on Victorian culture and society. In fact, generations of historians have also puzzled over this problem. Indeed, it would be hard to exaggerate the extent of scholarly interest in the social consequences of that suite of economic changes we label 'industrialisation'. As the academic discipline of history took shape in the early twentieth century, writers returned frequently to the theme of mechanisation and its (deleterious) impact on the labouring poor.[12] The emergence of a left-leaning social history in the post-war decades picked up this

historical problem and fed into one of the most influential history books of the twentieth century, E.P. Thompson's *The Making of the English Working Class*.[13] As social history bifurcated into economic history on the one hand and cultural history on the other in the following decades, the impact of the Industrial Revolution on living standards remained a bread-and-butter topic in British history and a focus for research by a variety of different methods and means – both quantitative[14] and qualitative.[15] As such, our own work fits into a rich vein of scholarly thinking about the social adaptation required by the coming of the machine age.

This book, however, is not concerned simply with tracing the parallels between these two historical moments of technological change. Rather, it seeks to bring them into conversation in much more active ways. What follows is a computationally intensive approach to the study of British society and culture in the era of industrialisation (c.1800–1920) bringing the technological innovations of AI and data science to the study of our past. The project underpinning the research for this volume is entitled Living with Machines. Our project shares its name with a pamphlet published in 1933 by the sociologist William Ogburn in which he advanced his theory of 'cultural lag': the contention that inventions outstrip societies' ability to understand and evaluate them. This is certainly the case for nineteenth-century Britain. One corollary of industrialisation in Great Britain was an explosion in the creation and collection of data. Initiatives such as the decennial population census (established 1801) and the Ordnance Survey (also 1801, though with earlier precedents), had their roots in the British state's response to war and the threat of invasion (it's called the 'ordnance' survey for a reason).[16] Over time, both evolved to become central planks of the emerging information state as they collected more and more detailed information about people and places.[17] They became a way of knowing, and potentially controlling, an increasingly complex and fluid civil society.

In tandem an 'information society' emerged, driven above all by the explosion in newspaper publishing across the nineteenth-century – itself the result of technological innovations reinforced by the elimination of punitive government taxes and rising living standards. In 1851 there were 365 newspapers in Britain, with London representing more than a third of the total; by the late 1880s that figure stood at over 2,000 separate titles, including 1,366 provincial newspapers.[18] In short, industrialisation produced a more complex society, one busily generating ever more records and documents about itself. By the century's end, industrial Britain

was creating vastly more information than individuals could realistically read, analyse and fully understand even in multiple lifetimes. This is the paradox that Ogborn identified, in which an increase of information goes in hand with a decrease of understanding.

We are now in a position to actively engage with this mass of data. Large swathes of the information curated by Victorian society and state has been digitised in recent decades, including household census returns from 1851, thousands of sheets of Ordnance Survey maps and billions of words from Victorian newspapers. Indeed, these digitised sources represent the bedrock of the present study. By shifting these documentary sources into machine-readable formats, libraries and archives have created the opportunity to know the past anew by mobilising the power of data science. The rapid pace of advancement in the field of AI and data-driven fields of research means that there is a genuine opportunity to ask a very different set of research questions, which could in turn be answered on a more empirically grounded and potentially more representative basis.

Despite the clear promise presented by the parallel availability of data and technology, so far only modest inroads have been made in this space. The reasons for this lag are complex and manifold. A major factor, however, has been the mismatch between the strengths and skills of the traditionally trained historian and those required to analyse and understand data at scale. There is of course a long history to data-driven history, which we touch on below. However, the 'two cultures' of the sciences and humanities identified by C.P. Snow in the late 1950s are still largely upheld by the drive to specialisation in the education systems of many countries, and the formation of disciplinary silos.[19] The result is that few historians today have the training to work at scale with the huge collections that the Victorians created. Another contributing factor, as we have discussed elsewhere, has been the need for libraries and archives to partner with commercial providers to fund their digitisation, due to lack of government investment, leading to data being siloed behind paywalls and limiting the extent of data processing that can be undertaken.[20] Taken together, the result has been another kind of 'cultural lag', meaning that there remains much to learn about the human, social and cultural consequences of this historical moment.

While access to data remains an ongoing challenge, one answer to the skills gap is collaboration. The Living with Machines project brought

together an unusually large and diverse team – data scientists, software engineers, historians, digital humanists, curators, library professionals, computational linguists, literary critics and an urban geographer – with the aim of undertaking research that at once offered a data-driven approach to history and a human-centred approach to data science. As such, the project was about *humanists* living with machines: working with computers, with methods from machine learning and with colleagues from computational backgrounds. This kind of exchange does not just happen but requires a proactive approach precisely because of the way that it brings together individuals from different research and professional cultures with different expectations about how to work, and how to disseminate research findings. In order that this labour not be rendered invisible, the project sought to share its experiences and recommendations in a short book, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project.*[21] One of the key principles that the authors draw attention to in those pages is that none of the represented groups or sectors should be in service of the others. Rather, we sought to create something that was greater than the disciplinary parts, building on the expertise, skills and experiences of the whole team.

The diverse makeup of the project team means that we have been able to make interventions back into the different fields from which our members have come, as well as those fields of discourse positioned at their interface – digital history and computational humanities. We have developed means to make the digitised collections of historical documents research-ready by processing them into more accessible formats; by releasing new contextualised datasets so that users can better understand the contours of what has been digitised; and by curating derived or sample datasets from behind paywalls to allow others to reproduce and build on our results. We have developed our own tools and methods for, among many other things: computationally 'reading' maps and identifying features such as rail and buildings (or anything else with a distinctive visual form); extracting place names (toponyms) from nineteenth-century texts and geolocating them; linking people across census returns; linking multiple datasets by location – such as maps, census returns, geolocated streets and stations – allowing for multidimensional analysis of changes to communities; and analysing the ways that that language evolved to describe new technological realities, specifically developing bespoke frameworks for annotating word meaning in historical texts, and new

algorithms for identifying changes in word meaning over time and across space. These contributions can be conceptualised as the foundations on which this book is built.

The decision to publish a monograph from this project – which was by no means a given – articulates our desire to have an impact on the historiography of the Industrial Revolution. It is aimed at historians (especially those who are open to the ways that data-driven approaches can contribute to the field), scholars practising digital and computational humanities, as well as researchers and practitioners in the GLAM (galleries, libraries, archives and museums) sector thinking about collections as data. The broader disciplinary and professional communities from which our team has drawn will perhaps engage with our book in a different way. The reason we chose to publish with the University of London Press and their instantiation of the Manifold platform was because we wished visibly to connect this capstone publication to all the undergirding forms of labour. The platform showcases some of these as additional resources that we developed: the datasets and digitised content; code, tools and software; data visualisations; and material we generated to make these materials more accessible, such as our video documentary series.[22] By bringing those together we hope to show what is now possible for other researchers who wish to use these resources, as well as driving credit back to all the different forms of labour that made this book possible (as discussed further below and in our authorship and credit statement). But the histories that we tell in the following pages are just a model of what it is possible to build from those foundations; we hope that others will develop these into larger, more ambitious projects in the future.

Writing the Industrial Revolution

While there are multiple contexts – intellectual, disciplinary, practical – for this book, our primary intellectual question is how people in the long nineteenth century learned to live with machines, which places our work within the more specific context of the accrued historiography surrounding the so-called 'Industrial Revolution'. What do historians mean when they talk of Britain undergoing the first 'Industrial Revolution', and what have been the main areas of debate over its timing, causes and consequences? It is in our contribution to this ongoing scholarship that we can assess whether the ways of working that we have championed in this book, and our wider

project, can truly extend our understanding of the changes wrought by the coming of the first machine age.

In the anglophone world, the popularisation of the concept of the 'industrial revolution' as a distinct epoch of social and economic transformation, beginning in Britain and subsequently spreading out across the world, is usually attributed to the economic historian Arnold Toynbee (1852–83). Toynbee had made a stir at Oxford with his lectures on Britain's industrial transformation, and, when he died suddenly, aged only thirty, his friends and colleagues published his notes and papers as *Lectures on the Industrial Revolution in England*.[23] Toynbee wasn't the first person to embrace the 'revolution' metaphor to dramatise industrialisation; broadly the same idea already had wide currency in Europe (not least through the greater awareness of the writings of Marx and Engels), but in Britain it only became an established feature of public discourse from the 1880s.[24] Toynbee intended the term to denote a very different type of 'revolution' to that advocated by Marx and Engels. He presented the 'industrial revolution' as the interconnected, economic corollary of the political revolution that popular Victorian 'Whig' historians had long associated with the restoration of Parliamentary sovereignty in the (supposedly) peaceful revolution of 1688. It was, in short, a key component of Britain's distinctive, non-revolutionary route to social transformation that should be contrasted with the violent, disorderly model of France in 1789.[25]

But no sooner had the expression begun to gain a degree of popular currency than the academic community began to question the existence of this supposed 'industrial revolution'. In the 1930s and 1940s a new generation of scholars began asking whether the Industrial Revolution was really so 'revolutionary', or indeed so located in 'industry', as Toynbee had confidently declared, and throughout the first half of the twentieth century the emphasis was firmly on the gradual rather than revolutionary nature of nineteenth-century economic change.[26] Of course, the writing of history rarely stands still. No sooner had scholars agreed that industrialisation proceeded in a gradual and piecemeal fashion than interpretative trends changed once again, and the fast-paced 'revolution' that Toynbee had postulated seemed to be back in vogue once more. The economic theorist Walt Rostow's 1960 book *The Stages of Economic Growth* provided a hugely influential model of industrialisation that formed a sharp contrast to the more gradualist models that had held sway in the interwar period. Here was an altogether punchier story, with the Industrial Revolution

marking a watershed in not just British but world history. For Rostow, industrialisation was the key to economic 'take off', and the promise of sustained economic growth.[27] But in little time, gradualism gained traction once more, at least for the British story, thanks to Nick Crafts, who in the 1980s recalculated economic growth rates for the period 1750–1830 and concluded that growth was much slower than Rostow's concept of rapid take-off permitted, sparking yet another round of revisionism concerning the timing, pace and nature of British industrialisation.[28] Given the paucity and incompleteness of economic data for this period, no definitive answer is possible, but it is important to recognise that even those most pessimistic about Britain's economic growth do not argue that little changed between the mid-eighteenth and mid-nineteenth centuries, that there was no 'industrial revolution'. On the contrary, their arguments are about why Britain experienced less dramatic improvements in productivity and gross domestic production than countries that industrialised later, *not* whether its economy and society underwent radical restructuring across these decades.[29]

For all the controversy surrounding the term 'industrial revolution', the concept has proved remarkably resilient, and for many decades there has been widespread agreement between academic historians that something happened in Britain in the years 1760–1860 that was so fundamental in nature that some form of special term is not only justified but needed. Where there has been less agreement is over the question of why Britain emerged as the first industrial nation. For much of the twentieth century, the answer seemed to lie in technology. Toynbee had made much of the role of inventors in shaping Britain's precocious industrialisation, and this emphasis on key inventions remains a staple of popular histories of the Industrial Revolution. Hargreaves' spinning jenny, Arkwright's spinning throstle and carding engine, Crompton's mule, Cartwright's power loom and James Watt's steam engine are regularly cited as the cause of Britain's elevation to its unique (though ultimately temporary) status as the world's industrial powerhouse.[30] In turn, this emphasis on inventive genius led historians to focus on those elements of British culture, education and political and legal frameworks that had supposedly predisposed its people to ingenious and entrepreneurial activity.[31]

As the title of our book, *Living with Machines*, indicates, we concur that the development and proliferation of new forms of technology in

the nineteenth century is an important historical event. But the complex forces that drove industrialisation should not be reduced to a simple narrative of technological determinism. Inevitably, new technologies played an important role in transforming Britain into an industrial society, but the world had already seen the emergence of some groundbreaking technologies – gunpowder, mechanical clocks, the printing press. None of these had ushered in an industrial revolution. Indeed, even Britain's leading sector – the textile industry – had mechanised long before the eighteenth century. Spinning had been partially mechanised by the invention of the spinning wheel in the fourteenth century; treadle looms were even older. Manufacturers had long sought to invent and improve machinery in order to find more efficient ways of creating goods that could be sold more cheaply and used more widely. These inventions served to enrich manufacturers and nations, but they had never previously amounted to anything that could be properly described as an 'industrial revolution'. The question remains, what changed in the eighteenth century?

Alongside technology, historians have drawn attention to many other possible causes of Britain's Industrial Revolution. Some point to the importance of agriculture in freeing labourers to work in industry;[32] of coal in powering the new industrial juggernaut and breaking the demographic constraints of an organic economy rooted in nature;[33] of the colonies in providing Britain with raw materials;[34] of slavery in providing the capital needed by industrialists;[35] of domestic industry and the labour provided by women and children;[36] and of rising demand for new goods.[37] Indeed, throughout the twentieth century, historians continued to argue about every aspect of the Industrial Revolution, producing in the end a complex, multilayered debate with very little agreement.

At the same time as some scholars argued over the origins and principal causes of the Industrial Revolution, so others argued over its consequences for people and, especially more recently, the environment. The question of the social impact of industrialisation was baked in from the time of Toynbee in the late nineteenth century.[38] Toynbee himself was a social reformer seeking to temper laissez-faire liberalism through an expanded role for an ameliorative state. Other early writers brought a similar social agenda to their historical enquiries about the social impact of industrialisation, building on a literary tradition that had long presented industrialisation as a source of immiseration.[39] In the twentieth century, economic historians such as T.S. Ashton pushed back against this thesis, insisting that

real wages rose substantially for the majority of the population during the Industrial Revolution.[40] The 'standard of living' debate increasingly focused on when real household incomes began to rise for the bulk of the population, and whether such monetary gains could be argued to have brought a better *quality of life* for the urban, industrial population. Some had no doubt not only that industrialisation brought higher real wages but that workers viewed such wages as providing adequate compensation for the 'urban disamenities' associated with industrialisation such as pollution, disease and, for many, reduced life expectancy.[41] Critics point out that adult male wage rates cannot tell the whole story, and that other data, such as per capita consumption trends, point to little or no improvement in working-class living standards before 1850.[42]

Again, there is no easy way to resolve these controversies. Much hinges on the question of how workers' themselves experienced the transition from agricultural to industrial production, and from rural to urban life. We know that Britain's emerging industries were highly regionalised, and that, in turn, wage rates and economic fortunes varied radically across the country. Bold statements about average living standards necessarily flatten this regional complexity, and ignore the extent to which old and new ways of living with machines existed side-by-side even in the same localities.[43] Social historians have sought to humanise the story that can be told through economic and demographic statistics, by turning to the accounts that working people themselves generated about the experience of industrialisation. E.P. Thompson's seminal book *The Making of the English Working Class* draws on a wide range of documentary and statistical sources, but at its heart is a sustained engagement with working-class radical print culture as an expression of the worldview of those oppressed and enslaved by the new industrial system.[44] Working-class autobiography has also proved a rich source for historians keen to restore working-class subjectivity to the history of the Industrial Revolution. The shadow of Thompson's study loomed large over early work in this vein, which tended to foreground radical politics and the 'birth and early struggles of the working class'.[45]

More recently politics and class-making have been decentred in favour of a greater emphasis on how industrialisation altered the fabric of working-class life and culture. Regenia Gagnier has focused specifically on shifting class-based models of selfhood in industrialising Britain, while more recently Emma Griffin has sought to recover how the social changes

associated with industrialisation could be understood as a form of personal liberation by many born into dependent poverty in the countryside.[46] Such accounts undoubtedly help to flesh out what industrialisation meant for the people who lived through it, but inevitably they, too, can be charged with flattening the diversity of human experience by privileging some voices over others. Did the working-class writers of radical pamphlets and newspapers really share the worldview of the men and women they claimed to speak for? Similarly, does reconstructing the subjectivity of a few hundred autobiographers mostly writing late in life allow us to make definitive statements about how most people experienced the Industrial Revolution? Sadly, there are no easy answers when our goal is to reconstruct the everyday experiences of people who have mostly left no written record beyond the official markers of birth, marriage and death.

Nor do we claim that *Living with Machines* can provide ready-made answers to issues which historians have argued over since the nineteenth century. Our claim is more modest, namely that the tools and methods developed here have the potential to open up a different perspective on the radical social transformations that everyone agrees remade British society between 1750 and 1900. We need to proceed with caution, constantly interrogating the provenance and partiality of the sources available for digital research. But as the chapters that follow demonstrate, there is now enormous potential to combine analogue and digital research methods to unlock facets of the everyday experience of industrialisation. Combining close and distant reading techniques can now be used to recover minority voices currently buried in the billions of words of digitised newsprint. Similarly, interconnecting historic maps, street-level census data and other sources of social and economic information, as we do in Chapter 3, can help us reconstruct everyday experience in all its complexity while still identifying meaningful, large-scale patterns. In short, *Living with Machines* is a first step on an important journey to bring history and computational methods together in new ways in order to help make sense of the past.

Digital history

Ours is by no means the first attempt to do data-driven history. Quantitative history has a long and rich lineage that takes in the emergence of 'cliometrics' in the late 1950s, and the pioneering demographic scholarship of the Cambridge Group for the History of Population and Social Structure, founded in 1964 by E.A. Wrigley and Peter Laslett. Our purpose here is

not to retell that story but to acknowledge that for many historians computational methods are still not the norm, even some eighty years later. At regular intervals there are calls to action, such as Jo Guldi and David Armitage's 2014 *History Manifesto*, which declared the need for a return to quantitative methods and *longue durée* history in response to the massive increase in the availability of data at scale, and Luke Blaxill's more recent intervention, calling on historians wedded to traditional methods to engage more directly with the radical challenge posed by digital methods.[47] Guldi and Armitage's manifesto, however, was met with a surprisingly fierce backlash. This probably owed less to the disciplinary clashes over cliometrics in the 1970s than to disciplinary suspicions that the tone of their manifesto echoed the grand claims for history as a social science that had been fashionable before the cultural turn of the 1990s. The great champions of historical sociology, such as Barrington Moore, Charles Tilly and Theda Skocpol, generally focused on the transformation of state and societal systems, rather than on quantitative methods, but their broad claims about historical processes have nonetheless often seemed at odds with the richly contextualised, humanistic and source-led methods of historians.

Indeed, much of the scepticism about digital history has stemmed from historians' suspicion that advocates of quantitative methods and big data have disregarded the need for painstaking source criticism. All too often, it has been easy to throw the old computer science adage of 'garbage in, garbage out' at headline-grabbing examples of what purports to be data-driven history. But there is a distinction to be made between humanities scholars engaging with data-driven methods and the kinds of sources and questions typically studied by historians and literary critics being 'scooped up by quantitative disciplines'.[48] An example of the latter is the now infamous 2011 'culturomics' article written by a team led by evolutionary biologists in cooperation with Google, which analysed millions of digitised books and introduced the Ngram viewer to the world.[49] Published in the high-profile journal *Science*, this article used the Google Books corpus to trace correlations between word usage frequency and known cultural trends (e.g. the usage frequency of the word 'slavery' in the corpus of books rose steadily across the nineteenth century peaking first during the American Civil War (1861–5) and again a century later during the civil rights movement (1955–68)). It also provided insights into the evolution of language (e.g. the rate of regularisation of verbs). However, the problems of the Google Books corpus for nuanced linguistic analysis are by now well

documented, including the impact of optical character recognition (OCR) errors, the over-representation of scientific literature, messy metadata, the equal weight assigned to each book regardless of its literary or commercial impact, and the compounded bias of aggregated source libraries.[50]

Despite greater understanding of the pitfalls of this dataset, it continues to be widely used in data-driven work. A 2019 article in *Nature Human Behaviour* used sentiment analysis on this corpus to plot historical well-being over more than two centuries in the United Kingdom, United States, Germany and Italy.[51] Perhaps seeking to address the well-rehearsed critiques, they 'corroborate' their results against alternative indices derived from independent corpora including the 'FindMyPast' corpus from the British Library's 'British Newspaper Archive', which we have also used on this project (although, as we discuss in Chapter 1, our aim has been to explore newspapers' hidden biases, rather than assume that problems of context and representativeness can simply be ignored if the corpus is large enough). The authors concluded that the data collected from books and newspapers is a reliable guide to the national mood because the plots correlated with key historic moments such as the American Civil War, the Wall Street Crash and Britain's 'Winter of Discontent' (1979). But bold claims such as the British were happiest in 1880 necessarily ignore key issues including the gradual democratisation of print culture in Victorian Britain and the emergence of a more sensationalist style of 'new journalism' at exactly this moment, as well as the changing meaning of words over the study's duration (while sentiment analysis dictionaries are commonly trained on recent online material). Such claims need an interrogation that the authors cannot give. Instead, they offer the caveat that 'Caution is needed when considering any long-run socioeconomic data. In all cases, there is a need for what historians call a "close read" of the historical data.'[52] But if a historian had been on the research team they would likely have done more than offer the data a 'close read'; they may have problematised the team's basic research paradigm, in the process probably undercutting the 'big find' status that has surrounded the reporting of this sort of data history project.[53]

The effect of such work is to rip digitised documents from their historical context. This, then, is not digital *history*. Echoing the rallying cry of Guldi in her recent book *The Dangerous Art of Text Mining*, we believe that there is a need to clarify and codify the distinction between good and bad digital methods in their application to historical data, and that

historians themselves must lead the way, rather than letting data scientists loose on this material in the absence of domain expertise and contextual knowledge.[54] One of the ways that we have sought to do this on Living with Machines was to develop new critical techniques and tools for contextualising and evaluating data at scale. This is the central goal of the digital 'Environmental Scan' method we developed for understanding hidden biases in large corpora of digitised newspapers.[55] But, as we discuss in Chapter 1, the same logic can and must be applied across digital methods in history. We would contend that the arguments for working in this way are not merely relevant to the needs and professional standards of historians trained to be attuned to the complexities, incompleteness and biases of their sources; they are also just good practice in data science, although still rarely undertaken in practice.

The balance we seek to strike here between working at scale and attending to historical contingency is related to the larger issue of how we talk about the payoff of digital approaches to history. Those working in this space for any time will know the strange expectations upon us. Colleagues are at once sceptical that computational approaches could tell them *anything* they didn't already know and scathing that these methods have not delivered a revolutionary new finding. The latter demand is perhaps due in part to the blustering early promises made for the potential of digital methods to transform knowledge, and deliver startling new insights.[56] But it is also due to the incursion of methods from scientific fields which tend to approach knowledge in a way that Wilhelm Windelband termed *nomothetic*, that is, having a tendency towards methods and approaches designed to derive generalisable laws to explain categories of phenomena. It is why historians have such problems with findings like those about 1880 being the year of peak 'happiness' in Britain. By contrast, scholarship in the arts and humanities is more comfortable with what Windelband calls the *idiographic* method, the tendency to specify what is contingent and particular.[57] Contributions to the field of knowledge in history tend, on the whole, to be incremental, detailed, grounded and alert to local inconsistencies, ambiguities and subjective experience.

In the collaborative research that we undertook for this book our process sought to bring together the idiographic and nomothetic in an iterative and complementary research process. Our aspiration is that digital approaches can allow us to generate statistical – or what some proponents of the digital humanities call 'distant' – overviews of our historical

documents and datasets; overviews which allow us both to determine the general trends and patterns within our data and to attend to the individual data points (determined by highly specific historical contexts) that make up the big picture. By keeping both worldviews in sight, and shifting between them as the research process unfolded, we have sought to define a particular research paradigm.

Victorian data

That research paradigm, however, is itself shaped by the historical records at our disposal in the digital age. As already indicated, the Victorians produced data on a scale beyond anything that can be read and analysed using the traditional historian's analogue methods. In the last few decades massive investments have been put into the digitisation of these records, which provide both new opportunities and new challenges for researchers. At the heart of this project are three key collections: newspapers, the population census and the Ordnance Survey maps, although these are surrounded by a constellation of other smaller datasets that we have accessed, digitised and shared (these are discussed more fully in the chapters that follow). This diversity of sources is key to our project; it offers a variety of perspectives and, crucially, novel ways of connecting these into a multidimensional picture of the past. Exploiting this diversity, however, requires an understanding of the provenance of digital data. After all, digital assets are not the same as the historical records themselves, either in their contours and coverage, or in their affordances; and, in turn, neither is a simple, unmediated window into past lives.

Provincial newspapers have long been recognised as a valuable and important historical source. In contrast to the national – in reality, metropolitan – press, local and regional newspapers put down deep roots in local society.[58] They contain numerous valuable micro-genres that speak to the events in people's lives, from news items and reports to family notices, letters from local communities and advertisements of the products that people might have had in their homes. Newspaper circulations grew enormously across the nineteenth century. Furthermore, as the content of newspapers was widely disseminated through reading rooms, clubs and through public readings, it was distributed more widely than circulation figures imply. That said, newspaper content was always slanted towards paying customers, and thanks to the influence of advertising often towards the worldview of the better off. Although in many

respects provincial newspapers represent a uniquely democratic source, this democracy had its limits. Such newspapers have long been mined by historians for the insights they can offer into all aspects of everyday life at the local scale, but their great value as a digital source is that they can also be used to aggregate such evidence at the regional or national scale.

The British Library's collection contains newspapers from 1603 to the present day, from both Britain and further afield. There are over 600,000 bound volumes of newspapers (occupying 32 kilometres, or 20 miles, of shelving) and over 300,000 reels of microfilm (occupying a further 13 kilometres, or 8 miles, of shelving). The size of this collection means that although digitisation of these newspapers began over two decades ago, it is still an ongoing process. The data we are working with derives from a range of different projects to digitise the British Library's newspapers. The largest is the British Newspaper Archive (BNA), which is a collaboration between the British Library and the genealogy website FindMyPast, which offers access to the digitised content mainly through a subscription service. In a typical week in June 2023, 23,182 new pages were added to this database.[59] The project worked with a version of this dataset that comprises 68 million pages with 150 billion words from circa 1,600 newspaper titles spanning 1780–1920. We are the first project to have been provided access to (large swathes of) this important dataset.[60]

The FindMyPast data also subsumes important earlier digitisation efforts resulting from the British Library's prior collaborations first with JISC (formerly the Joint Information Systems Committee) and then with digital publisher Gale. In addition, our project brought in new digitisations of British Library holdings undertaken by the Library's Heritage Made Digital (HMD) project, including a significant tranche funded directly by the project. HMD's newspaper digitisation programme has been designed specifically to prioritise the digitisation of rare and vulnerable titles deemed to be of historical importance. In turn, our own digitisation selection prioritised longer-running titles from industrial districts, especially those known to have sought a working-class readership.[61] But despite the mammoth scale of newspaper digitisation, it is important to remember that far more newspapers remain undigitised, probably more than 80 per cent, some of which are too poorly preserved to be digitised – and other historic newspapers have not survived at all. Moreover, digitisation has neither progressed in a completely systematic nor a completely random way. Rather, successive waves of digitisation have each been shaped by the priorities of

the organisations that led them. JISC and Gale both focused on research users (principally working via search portals). They prioritised quality provincial newspapers with long print runs: newspapers that could make a plausible claim to have been serious newspapers of record. By contrast, FindMyPast's target audience is the general public, or rather that segment of it interested in family and local history. Their selection criteria are not a matter of public record, but they will certainly have been different.

The digital 'Environmental Scan' is our answer to the problem of how to know the relationship between what has, and has not been, digitised, and the potential implications of this for digital research. It is a tool that enables researchers to conduct source criticism at scale. It does this by deploying information derived from contemporary reference works to help us better understand the composition of the big bag of words that is any large text corpus. For newspapers, we use information extracted from the Mitchell's Press Directories, established in 1846 and published annually from 1856 (these were digitised and structured within the project, for details see Chapter 1). We use the information derived from the digitised Press Directories to map the approximate profile of the newspaper publishing environment between the 1840s and 1920 (our end point), and then compare this profile with that of different newspaper corpora (JISC, Gale, BNA, etc.) to understand how selection decisions have influenced the representativeness of the different digital samples. The Environmental Scan is thus a tool for understanding bias in digital sources, but we are careful not to argue that there is some hypothetical 'ideal' sample that would eliminate all bias. Firstly, although we have no robust information about newspaper circulation after 1856, it is evident that not all newspapers were equal in their importance, influence or even their ability to represent facets of contemporary public discourse. JISC and Gale had good reason to favour some newspapers over others when spending scarce resources. Secondly, newspapers as a whole were (and are) an imperfect reflection of contemporaneous language precisely because they were mediated by commercial imperatives that gave some voices, and worldviews, more prominence than others. The Environmental Scan simply makes it possible for the first time to map these inclusions and exclusions more precisely.

Alongside the newspaper archive, the project also works with the census. The census began in 1801 but was initially little more than a basic headcount. The first full household census was completed in 1841, and included information about the occupation, place of birth and relationship

to the head of household of eighteen million people living in Britain at that time. Regrettably, the original household returns from 1841 were destroyed, and we only have aggregated information about the population, but from 1851 the decennial census returns survive as raw data, as well as in aggregated form. Census returns from across the country were transcribed by hand by government officials and bound into the Census Enumerator Books (CEBs), which were subsequently stored at the Public Record Office, now the National Archives.[62] The provenance of the census is clearly very different from that of the newspapers. Whereas the latter were produced, for commercial reasons, by numerous different individuals within society, the census was funded and produced by the Victorian state with a view to understanding the size and structure of its citizenry and economy. As a decennial snapshot of households (both their demographic composition and their economic underpinning), the census has long been exploited by academic researchers. It is of interest to our project primarily for the detail it gathered about occupations, which in turn provides a window on to new technologies invented in the nineteenth century and the ways in which they proliferated into, and impacted, the lives of ordinary workers.

It was not questions of an academic nature, however, that underpinned the original digitisation of the census. By the twentieth century, the census had come to the attention of a growing community of family historians and so genealogy companies, including (once again) FindMyPast, took an early interest in transcribing the original CEBs for the lucrative family history market. These transcriptions, produced with the family historian in mind, were not suitable for academic research. However, a subsequent ESRC-funded project at the University of Essex cleaned, standardised and coded FindMyPast's transcriptions to produce a structured dataset of the kind required by researchers. It is known as Integrated Census Microdata (I-CeM).[63] I-CeM contains the records of over 180 million individuals enumerated in British censuses between 1851 and 1911, including over 100 variables for each (I-CeM lacks data for England and Wales in 1871 and Scotland in 1911). The I-CeM dataset enables academic research on the British census in a way that FindMyPast's original digitisation did not, and already forms the basis of a large and growing literature in the economic history tradition exploring aspects of family, demography, occupations and migration in nineteenth-century Britain.[64]

While the census provides us a point of entry into the living and working conditions of the people who experienced the Industrial Revolution,

understanding the impact of industrialisation on the environment they inhabited is harder to capture. Our closest proxy, perhaps, is the series of Ordnance Survey maps. This mapping project was begun at the turn of the nineteenth century as a government-led defence project in response to threats from France to England's Kentish border. Once again, what gets mapped, and how it is represented, are political issues. Maps, like newspapers and population censuses are human documents, socially produced with specific ends in mind. Equally, however, maps can be interrogated with questions about the physical environment and its transformation in mind. They provide a powerful visual record of the natural and built environment, and – potentially – of changes reshaping that landscape, given that places were re-surveyed over time for revised and new editions of map series at different scales. Maps are of course being scanned by the thousands at cultural heritage institutions around the world. Our project has drawn heavily on the work of the National Library of Scotland (NLS). Since 2014, the NLS has been able to scan over 20,000 sheet maps per annum, and now has over 250,000 maps available online. The NLS has made those images freely viewable online, and the associated geo-data available for non-commercial use according to the principles of open and reproducible research. We supplemented the gaps in the NLS's digitised Ordnance Survey map coverage with new digitisation from copies of the maps held in the British Library's collection. Our work to date has chiefly focused on the second edition County Series six-inch to one mile maps (1888–1914), but in the spirit of the Environmental Scan our early work on the maps sought to understand how the surveying process played out over time, and the impact that temporal process has on the way that we ask questions of such a dataset.[65]

The project therefore brings together multiple very large collections that present different forms of data: textual, tabular and visual. However, despite the size, breadth and complexity of the datasets, it is important to pause a moment to recognise the narrow world from which they are mostly drawn, and the limitations they therefore present. By their nature these sources tend to accentuate the domestic, white mother-country facets of the story of Britain's industrialisation, largely obscuring the centrality of empire, slavery and global trade in general to Britain's emergence as an economic and imperial superpower in this period. Entanglement with the empire ran through all aspects of daily life in nineteenth-century Britain, from the mass consumption of imported tea and sugar to the reliance of

so many of its industries on imperial raw materials and markets and the trading advantages secured through British naval power. Both Ordnance Survey maps and the decennial census were confined within the geographical and political boundaries of the nation state; Ireland was included, the empire overseas was not (though as core technologies of state power both featured as tools of colonial rule).[66]

These sources still register traces of the wider story, but they are fleeting and partial. For example, the census recorded 'place of birth', including when people were born overseas; maps captured place names that encode local connections with empire and global trade, as well as recording the physical infrastructure that made that trade possible; and some newspapers offered reports not just of world events but also of other international developments that might affect the local economy (such as accounts of harvests, shipping, mineral discoveries or overseas infrastructure projects). Newspapers can also tell us about people's daily engagement with commodities sourced from the empire. However, the ways in which such traces are refracted through the interests and the default worldview of imperial power mean that historians need to read 'against the grain' of such documents. As such, most research in this space has relied on painstaking archival work, supplemented by clues drawn from local and colonial newspapers, to tell alternative, de-centred stories of industrialisation.

Of course, the work needed to uncover alternative histories capable of decentring the metropole need not be separate from computational research. Indeed, the archival work described above can build up new digital resources over time, even if they are on a different order of magnitude.[67] And digital methods can be developed to scale up critical modes of close reading (focusing on discovery of rare or occluded phenomena, rather than large-scale patterns). However, the emphasis of this current project on what is possible *already* with these large and canonical datasets means that this volume often reproduces the dominant domestic narrative. Nevertheless, we do seek in the following pages to identify how our tools and methods could be extended to other collections and materials, and how we might be able to further develop computational approaches to find those subtle signals that can help us to capture an alternate worldview.

Working with machines

The foregoing summary of our data begins to suggest the breadth of skills required of a team seeking to tackle such sources, both in terms of the

domain expertise in the specific historical documents and the kinds of computational approaches needed to leverage such data types at scale. Our assembled team featured computational linguists for analysing our large-scale text corpora, researchers able to work with regular expressions for the census, and colleagues conversant with Geographical Information System (GIS) and computer vision for the maps; our project as a whole was underpinned by the expertise of data scientists and research software engineers. However, assembling diverse skills does not necessarily lead to interdisciplinary research. The risk remains that people will retreat into their received ways of working. This leads to multidisciplinary, rather than interdisciplinary, work: it may draw on knowledge from different disciplines, but research and its outcomes ultimately stay within traditional boundaries. By contrast, the decision to work in a fully interdisciplinary fashion has ramifications both for the research process and its outcomes – as well as to whom such outcomes speak. We sought to find collaborative agendas which drew expertise from multiple fields in order to develop research questions and methods for answering these. In this way we challenged each member of our team to work within a new and unfamiliar research paradigm.

The combination of skills brought together by the authors of this book, and the broader team that sits behind it, was enabled by the vision of two institutions. The Living with Machines project began in 2017 as a conversation between the British Library and the Alan Turing Institute based on the clear opportunities their co-location on the British Library's main St Pancras site posed for collaboration. The Alan Turing Institute is the UK's national institute for data science and AI, with strong support from government research funders and the university sector. Together, the two institutions provided a complementary set of offerings: the British Library has digitised millions of pages from its collections and has amassed unparalleled expertise in both the curation and dissemination of cultural heritage, while the Alan Turing Institute has the expertise to harness cultural heritage data to answer research questions at scale. Nonetheless, the collaboration was experimental in nature, since neither institution had previously attempted a cross-disciplinary collaboration so radical in nature.

The project was designed with both general and specific aims. The undergirding aim of the project was to develop new computational techniques to marshal the UK's rich historical collections so that new research questions could be posed. This aim was intentionally broad: the project

was intended to explore the research rendered possible by more than two decades of digitisation and investment in data creation. The digitised collections of the long nineteenth century were proposed as the test bed for this more general ambition, due in part to the British Library's long-term investment in the digitisation of its newspapers. Our more specific aim was to provide new perspectives on the effects of the coming of the machine age on people's everyday lives. However, history is only one of the fields in which we hoped to intervene – a fact not necessarily apparent when encountering the text of this book in isolation. More broadly, we aspired to generate outcomes not only back into the constituent disciplines represented by our large team but also those interdisciplinary spaces at their interstices. We wanted to develop generalisable tools, code, and infrastructure that could be adapted for, and help inspire, future research projects. Crucially, we aimed not only to feed through developments from the computational sciences into humanities research applications but also to reverse this trend by generating new methods from a humanities project that could make interventions within the fields of data science and other computational research. And we have been successful in doing this, as the example of our MapReader software (discussed in Chapter 2) demonstrates: while this computer vision pipeline was imagined in the first instance as a tool for the 'distant viewing' of historic maps, it was also taken up by the SciVision project at the Turing, for analysing plant phenotypes.[68]

One of the ways that Living with Machines is marked apart from digital projects from as little as ten years ago is its emphasis on *analysis*. To generalise somewhat: the previous generation of projects were more often focused around data collection, and/or the curation of data into queryable databases. Our ambition was to use previously created resources and datasets to drive applied analysis, thereby moving past the 'perpetual future tense of digital history'.[69] The larger team were distributed across a set of broad research agendas, which were launched initially through project labs designed to stimulate experimentation: on the language of mechanisation, on changes that can be mapped in time and space, on the bias and representation of our sources, on community engagement and on tools and infrastructure development. Each lab assembled a team with members drawn from different disciplines and professional backgrounds, and their first task was to design an initial research task (what we termed the 'minimum research outcome') that spoke back to the range of represented disciplines.[70] We wanted work

that answered a concrete historical research question, that provided an opportunity to explore the question with methods from data science and that identified a specific data set that it could leverage. Outcomes from this work could be shared as code, software or pipelines for preparing and analysing data, shared in code repositories for the benefit of others working with cultural heritage data; as new datasets (or old datasets rendered newly research-ready) published on repositories and described within data papers; as articles reporting new methods or findings, which might speak to communities in the fields of computational linguistics, GIS, computer vision or digital history, among others; as well as historical research outcomes taking the more traditional form of articles and books; or as all of the above.

This may sound abstract. This is partly because our research proceeded differently for each of the sub-teams: the group tasked with analysing how the language used to describe the mechanisation of work and life evolved over the nineteenth century worked in very different ways from the sub-team that developed methods to quantitatively describe the biases of digitised newspaper collections. Nevertheless, to grasp what interdisciplinary collaboration might look like in action, it is perhaps useful to describe the specific research process underpinning one of our chapters: Chapter 3, 'Beyond the Tracks'. This chapter reconstructs the physical impact of rail infrastructure on the lived environment, both urban and rural, in order to explore residential patterns that could better illuminate Victorian perceptions of the contrasting benefits and disbenefits of living close to rail. Who sought to live close to a railway station, and to what extent did wealth allow some to maximise the network-amenity benefits of the rail system while minimising their exposure to the noise, pollution and noxious trades often associated with proximity to dense rail infrastructure? To answer this question we needed to know the footprint of rail and its surrounding infrastructure on the ground (data we derived from maps), the precise location of passenger railway stations, and where people from different backgrounds lived at the level of the street rather than larger administrative areas (census enumerators' data, cross referred with datasets that could geolocate this data at the street level).

The analyses for this chapter relied heavily on the contribution of postdoctoral researchers with extensive prior experience using I-CeM, the computerised dataset of Victorian and Edwardian census enumerators' returns described above. Joshua Rhodes played the lead role on the

chapter, developing a new method for geo-linking census street addresses using only open-source spatial sources (GB1900 and OS Open Roads),[71] by locating individuals and households from the census (along with all the encoded socio-economic data about them in I-CeM) more precisely on the ground. This is a major feat of data curation in its own right and can unlock numerous different research questions beyond this specific application.

The 'Beyond the Tracks' chapter took the form it did because other postdoctoral researchers and data scientists on the project had already put a great deal of work into developing research tools designed to facilitate new forms of spatial research. Historians Katherine McDonough and Daniel C.S. Wilson worked with Kaspar Beelen and Daniel van Strien, a digital humanist and digital curator respectively, in collaboration with Kasra Hosseini, a research software engineer, to develop the aforementioned computer vision tool, MapReader, which harnesses machine learning to automatically identify distinctive features such as buildings and railway infrastructure on digitised Ordnance Survey maps (after manual training).[72] It was this research that allowed us to identify the physical footprint of the rail network ('railspace') across the whole country, and therefore to measure the density of rail infrastructure at the micro-level. This is very different from more conventional GIS methods that focus exclusively on the network-amenity dimension of railways by reconstructing them as vectors connecting points on a map, rather than as physical entities transforming people's lived environments in very uneven ways depending on the density of rail and its impact on residential space. Separately, Mariona Coll Ardanuy, Giorgia Tolfo and others worked to convert an important account of historic passenger railway stations, Michael Quick's *Railway Passenger Stations in Great Britain: A Chronology*,[73] into a machine-readable format, which formed the basis of our new dataset, StopsGB: Structured Timeline of Passenger Stations in Great Britain.[74]

The 'Beyond the Tracks' chapter was therefore conceived as a conscious attempt to bring this collection of individuals, skills and new tools into dialogue with each other to ask how far late Victorian residential patterns reflected the influence of rail as both amenity and disamenity. Did workers in certain occupational groups tend to live close to passenger stations? Were indicators of wealth such as live-in servant-keeping and occupation associated with a disinclination to live close to very dense rail infrastructure? Who did live closest to dense rail, and were these workers in the poorest groups, or perhaps those associated with trades that tended to be

located close by dense rail? For a long time this research was known as the 'convergence experiment', a term which in many ways embodies the ethos of the project, and the way in which it sought to bring together not just very different types of sources but also researchers with very different skill sets. Some of the key findings from this experiment are discussed in Chapter 3.

This language of convergence is also important for thinking about what it means to credit all the labour and expertise that goes into developing such research, with all its varied inputs and dependencies. We have sought to ensure that credit flows to all involved in the work of a chapter – at all stages from data preparation and conceptualisation, through to writing – by implementing a capacious and generous model of authorship, as we discuss further in our authorship and credit statement. The long author lists show the breadth of expertise brought to bear on each chapter; and the recurrence of contributor names across multiple chapters show the ways that different interests create threads of connection between the chapters. Those threads are evidence of that collaboration across the project, as well as within more discrete experiments. But the list of the chapter authors alone fails to capture the broader ecosystem of labour on the project that sits outside the unit of the chapter. This is one of the reasons that we chose to publish this book using University of London Press's Manifold platform, as we touched on briefly above. The way that the platform can host supplementary resources allows us to show the reader our 'workings'. As well us allowing us to link to the data, code and other resources underlying each chapter, the Manifold platform has enabled us to draw together resource collections that represent the other substantive outcomes of Living with Machines including: other publications; datasets and new digitised content; software, tools, pipelines and models; videos from the Living with Machines documentary series; and data visualisations. We thus make visible the kinds of labour that do not result in authorship.

Moreover, we believe that by situating our book in this broader context we enable readers from a range of backgrounds to make use of it in a way that meets their interests and needs. The reader can choose simply to follow the text with its figures and tables or navigate out to interrogate the underlying data or code. They can watch our docu-series videos as introductions to the key areas we have worked in on the project or seek deeper understanding of the methodological underpinnings of some of the chapters by reading other publications on that topic. As such, we hope that this format opens up conversation and exchange with the broader

community, rather than positioning the books as the final word on the topic. We have sought to foster that open conversation by sharing drafts of the chapters in open review, allowing our interlocutors to see our work in progress and to comment on it. We have been struck by the great generosity and collaborative spirit of the panel of reviewers, who have helped us to improve the book. Even though the texts of the chapters themselves are now fixed, we hope that the conversation will continue. For us, the best response we could hope for would be if our readers reuse our data, methods and tools, to respond to the provocations we make in the following pages, or – perhaps even better – to tackle challenges in other fields and domains.

Notes

1 Ahnert acted as lead author; all authors contributed equally to the writing to the revision process.

2 For instance, 'Goldman Sachs Report: ChatGPT Could Impact 300 Million Jobs', *Open Data Science*, 5 April 2023, accessed 23 February 2026, https://odsc.medium.com/goldman-sachs-report-chatgpt-could-impact-300-million-jobs-4f375b30e606; 'Pause Giant AI Experiments: An Open Letter', futureoflife.org, 22 March 2023, accessed 23 February 2026, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

3 J.I. Gershuny and R.E. Pahl, 'Work Outside Employment', *New Universities Quarterly* 34 no. 1 (1979): 120–35, and J.I. Gershuny and R.E. Pahl, 'Britain in the Decade of the Three Economies', *New Society* 3 (1980): 7–9.

4 Jefferson Cowie, *Stayin' Alive: The 1970s and the Last Days of the Working Class* (The New Press, 2010); Sherry Lee Linkon, *The Half-Life of Deindustrialization: Working-Class Writing about Economic Restructuring* (University of Michigan Press, 2018); Lutz Raphael, *Beyond Coal and Steel: A Social History of Western Europe after the Boom*, trans. Kate Tranter (Polity Press, 2023).

5 For a sober analysis of the moment, see Emily M. Bender, 'Policy Makers: Please Don't Fall for the Distractions of #AIhype', *Medium*, 29 March 2023, accessed 23 February 2026, https://medium.com/@emilymenonbender/policy-makers-please-dont-fall-for-the-distractions-of-aihype-e03fa80ddbf1.

6 Klaus Schwab, 'The Fourth Industrial Revolution: What It Means, How to Respond', *World Economic Forum*, 14 January 2016, accessed 23 February 2026, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

7 The term entered the English language by convoluted means. Early uses can be identified in France where it was conceived as 'la révolution industrielle' as a counterpoint to the political revolution that had occurred in France (for example, Natalis Briavoinne, *De l'industrie en Belgique: Sa situation actuelle*

– *Causes de décadence et de prospérité* (A. Wahlen, 1839), 185–210) and Germany (Friedrich Engels, *The Condition of the Working Class in England*, ed., with introduction, David McLellan (Oxford University Press, 1993), 18). It entered the lexicon when Engels' work was translated in the late nineteenth century.

8 Thomas Carlyle, 'Sign of the Times', *Edinburgh Review* (1829): 61.

9 William Wordsworth, 'Outrage Done to Nature', from *The Excursion* (Longman, Hurst, Rees, Orme, and Brown, 1814); William Blake, 'And Did Those Feet in Ancient Time', from *Milton, A Poem*, ed. Robert N. Essick and Joseph Viscomi (Princeton University Press, 1993).

10 Benjamin Disraeli, *Sybil, Or the Two Nations* (Henry Colburn, 1845); Elizabeth Gaskell, *Mary Barton* (Chapman & Hall, 1848); Charles Dickens, *Hard Times* (Bradbury & Evans, 1854); George Gissing, *Demos* (Smith, Elder & Co., 1886).

11 John Stuart Mill, *Principles of Political Economy*, ed. W.J. Ashley (Longmans, Green and Co., 1909), 751. For the later nineteenth century, see J.A. Hobson, 'The Influence of Machinery upon Employment', *Political Science Quarterly* 8, no. 1 (1893): 97–123; J. Shield Nicholson, *The Effects of Machinery on Wages* (Swan Sonnenschein & Co., 1892).

12 John H. Clapham, *An Economic History of Modern Britain: The Early Railway Age, 1820–1850* (Cambridge University Press, 1926); J.L. Hammond and Barbara Hammond, *The Skilled Labourer* (Longmans, Green and Co., 1919).

13 E.P. Thompson, *The Making of the English Working Class* (Penguin Books, 1976).

14 N.F.R. Crafts, *British Economic Growth during the Industrial Revolution* (Clarendon Press, 1985).

15 Herbert L. Sussman, *Victorians and the Machine: The Literary Response to Technology* (Harvard University Press, 1968); B. Rieger, *Technology and the Culture of Modernity in Britain and Germany, 1890–1945* (Cambridge University Press, 2005); Tamara Ketabgian, *The Lives of Machines: The Industrial Imaginary in Victorian Literature and Culture* (University of Michigan Press, 2011); Wolfgang Schivelbusch, *The Railway Journey: The Industrialization of Time and Space in the Nineteenth Century* (University of California Press, 2014).

16 James Vernon, *Distant Strangers: How Britain Became Modern* (University of California Press, 2014).

17 Mary Poovey, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (University of Chicago Press, 1998); Patrick Joyce, *The State of Freedom: A Social History of the British State since 1800* (Cambridge University Press, 2013). For a more sceptical assessment of the knowing state see Edward Higgs, 'The Rise of the Information State: The Development of Central State Surveillance of the Citizen in England, 1500–2000', *Journal of the History of Sociology* 14, no. 2 (2001): 175–97. For international and pre-modern comparisons, see also Bruce Curtis, *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875* (University of Toronto Press, 2002); and Jacqueline Wernimont, *Numbered Lives: Life and Death in Quantum Media* (MIT Press, 2019).

18 Alan J. Lee, *The Origins of the Popular Press, 1855–1914* (Croom Helm, 1976), 290–1.

19 See C.P. Snow, *The Two Cultures, with an Introduction by Stefan Collini* (Cambridge University Press, 1998).

20 Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009175548.

21 Ahnert et al., *Collaborative Historical Research*.

22 https://read.uolpress.co.uk/projects/living-with-machines.

23 Arnold Toynbee, *Lectures on the Industrial Revolution in England; Popular Addresses, Notes and Other Fragments together with a Short Memoir by B. Jowett* (Rivingtons, 1884).

24 A simple Google Ngram search confined to British English sources records sees an increase in occurrences from c.1880 with a continuous increase in the frequency of usage from 1901 to 1944.

25 Alon Kadish, 'Arnold Toynbee (1852–1883)' *ODNB* [2004].

26 Joseph Schumpeter, *Business Cycles*, vol. 1 (McGraw-Hill Book Company, 1939), 67–8; Clapham, *Economic History*, 143–5; E.M. Carus-Wilson, 'An Industrial Revolution of the Thirteenth Century', *Economic History Review* 11 (1941): 39–60; J.U. Nef, *The Rise of the British Coal Industry*, vol. 1 (George Routledge & Sons, 1932), 165–89; J.U. Nef, 'The Progress of Technology and the Growth of Large-Scale Industry in Great Britain, 1540–1640', *Economic History Review* 5, no. 1 (1934): 24.

27 Walt Whitman Rostow, *The Stages of Economic Growth: A Non-Communist Manifesto* (Cambridge University Press, 1960). See also Phyllis Deane, *The First Industrial Revolution* (Cambridge University Press, 1967), 117; Phyllis Deane and W.A. Cole, *British Economic Growth, 1688–1959: Trends and Structure* (Cambridge University Press, 1962).

28 Crafts, *British Economic Growth*.

29 For a useful summary of these arguments, see M.J. Daunton, *Progress and Poverty: An Economic and Social History of Britain 1700–1850* (Oxford University Press, 1995), 125–36.

30 For instance, Historyhit.com, https://www.historyhit.com/key-figures-of-the-british-industrial-revolution/; schoolshistory.org, https://schoolshistory.org.uk/topics/british-history/industrial-revolution/inventors-and-inventions/.

31 For an example of such an approach, see Eric Jones, *The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia*, 3rd ed. (Cambridge University Press, 2003).

32 Patrick O'Brien, 'Path Dependency, or Why Britain Became an Industrialized and Urbanized Economy Long before France', *Economic History Review* 49, no. 2 (1996): 213–49; R.C. Allen, *The British Industrial Revolution in Global Perspective* (Cambridge University Press, 2009), chap. 3.

33 E.A. Wrigley, *Continuity, Chance and Change: The Character of the Industrial Revolution in England* (Cambridge University Press, 1988).

34 Ronald Finlay and Kevin H. O'Rourke, *Power and Plenty: Trade, War, and the World Economy in the Second Millennium* (Princeton University Press, 2008).

35 Eric Williams, *Capitalism and the Slave Trade* (University of North Carolina Press, 1944); Joseph E. Inikori, *Africans and the Industrial Revolution in England: A Study in International Trade and Economic Development* (Cambridge University Press, 2002); Sven Beckert, *Empire of Cotton: A Global History* (Alfred A. Knopf, 2014).

36 Maxine Berg, 'What Difference Did Women's Work Make to the Industrial Revolution?', *History Workshop Journal* 35, no. 1 (1993): 22–44.

37 Jan de Vries, *The Industrious Revolution: Consumer Behaviour and the Household Economy* (Cambridge University Press, 2008); Maxine Berg, *Luxury and Pleasure in Eighteenth-Century Britain* (Oxford University Press, 2005).

38 Toynbee, *Lectures*.

39 Sidney Webb and Beatrice Webb, *Industrial Democracy*, 2 vols (Longmans, Green and Co., 1898); Hammond and Hammond, *The Skilled Labourer*.

40 T.S. Ashton, *The Industrial Revolution* (Oxford University Press, 1948).

41 J.G. Williamson, 'Was the Industrial Revolution Worth It? Disamenities and Death in Nineteenth-Century British Towns', *Explorations in Economic History* 19, no. 3 (1982): 221–45.

42 Sara Horrell and Jane Humphries, 'Old Questions, New Data, and Alternative Perspectives: Families' Living Standards in the Industrial Revolution', *Journal of Economic History* 52, no. 4 (1992): 849–80; Joel Mokyr, 'Is There Still Life in the Pessimist Case? Consumption during the Industrial Revolution', *Journal of Economic History* 48, no. 1 (1988): 69–92.

43 Pat Hudson, ed. *Regions and Industries: A Perspective on the Industrial Revolution in Britain* (Cambridge University Press, 1989); Maxine Berg and Pat Hudson, 'Rehabilitating the Industrial Revolution', *Economic History Review* 45 (1992): 24–50.

44 Thompson, *English Working Class*.

45 David Vincent, *Bread, Knowledge and Freedom: A Study of Nineteenth-Century Working Class Autobiography* (Europa Publications, 1981), 3.

46 Regenia Gagnier, *Subjectivities: A History of Self-representation in Britain, 1832–1920* (Oxford University Press, 1991); Emma Griffin, *Liberty's Dawn: A People's History of the Industrial Revolution* (Yale University Press, 2013). See also Jamie L. Bronstein, *The Happiness of the British Working Class* (Stanford University Press, 2023).

47 Luke Blaxill, 'Why Do Historians Ignore Digital Analysis? Bring on the Luddites', *Political Quarterly* 94, no. 2 (2023): 279–89.

48 Ted Underwood, 'Dear Humanists: Fear Not the Digital Revolution Advances in Computing Will Benefit Traditional Scholarship – Not Compete with It', *The Chronicle of Higher Education*, 27 March 2019, accessed 23 February 2026, https://www.chronicle.com/article/dear-humanists-fear-not-the-digital-revolution/.

49 Jean-Baptiste Michel et al., 'Quantitative Analysis of Culture using Millions of Digitised Books', *Science* 331 (2011): 176–82.

50 E.A. Pechenick, C.M. Danforth and P.S. Dodds, 'Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-cultural and Linguistic Evolution', *PLoS ONE* 10, no. 10 (2015): e0137041.

51 Thomas T. Hills, Eugenio Proto, Daniel Sgroi and Chanuki Illushka Seresinhe, 'Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books', *Nature Human Behaviour* 3, no. 12 (2019): 1271–5.

52 Hills et al., 'Historical Analysis of National Subjective Wellbeing', 1274.

53 'Scientists Pinpoint the Year That Britons Were Happiest', *Independent*, 15 October 2019; 'Victorian Times Were Happiest, Study of National Mood Finds', *The Times*, 14 October 2019.

54 See Jo Guldi, *The Dangerous Art of Text Mining: A Methodology for Digital History* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009263016.

55 Kaspar Beelen, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan', *Digital Scholarship in the Humanities* 38, no. 1 (2023): 1–22, https://doi.org/10.1093/llc/fqac037; Kaspar Beelen, Jon Lawrence, Katherine McDonough and Daniel C.S. Wilson, 'Whose News? Critical Methods for Assessing Bias in Large Historical Datasets', *Computational Humanities Research* (2025): 1–21.

56 For an account, see Cameron Blevins, 'Digital History's Perpetual Future Tense', in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016), https://doi.org/10.5749/9781452963761.

57 Wilhelm Windelband, 'History and Natural Science', *Theory & Psychology* 8, no. 1 (1998): 5–22.

58 Andrew Hobbs, *A Fleet Street in Every Town: The Provincial Press in England, 1855–1900* (Open Book Publishers, 2018).

59 Statistics true as of 21 June 2023.

60 Although some prior publications used this data, analysis was done on the FindMyPast servers, and the researchers were not given direct access to the data assets. See for example Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper team and Nello Cristianini, 'Content Analysis of 150 Years of British Periodicals', *Proceedings of the National Academy of Sciences* 114, no. 4 (2017): e457–65. The data underpinning this work was published as n-grams: Saatviga Sudhahar,

Nello Cristianini, Thomas Lansdall-Welfare and the FindMyPast Newspaper team, 'FindMyPast Yearly N-grams and Entities Dataset', 2016, https://doi.org/10.5523/bris.dobuvuu00mh51q773bo8ybkdz.

61 Giorgia Tolfo et al., 'Hunting for Treasure: Living with Machines and the British Library Collection', in *Digitised Newspapers: A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*, ed. Estelle Bunout, Maud Ehrman and Frédéric Clavert (De Gruyter, 2023), https://doi.org/10.1515/9783110729214-002.

62 Michael Anderson, *The 1851 Census: A Guide to the National Sample of the Enumerators' Returns* (Cambridge University Press, 1987). See also Ahnert et al., *Collaborative Historical Research*.

63 I-CeM has been available since 2014 and was updated in 2024 alongside the release of the 1921 census microdata. Kevin Schürer and Edward Higgs, *Integrated Census Microdata (I-CeM), 1851–1911* [data collection], UK Data Service, SN: 7481, 2022, https://doi.org/10.5255/UKDA-SN-7481-2; K. Schürer and E. Higgs, *Integrated Census Microdata (I-CeM) Names and Addresses, 1851–1911: Special Licence Access* [data collection], 2nd ed., UK Data Service, SN: 7856, 2022, https://doi.org/10.5255/UKDA-SN-7856-2; K. Schürer and A. Wakelam, *Integrated Census Microdata (I-CeM), England and Wales, 1921* (2024) [data collection], UK Data Service, SN: 9280, https://doi.org/http://doi.org/10.5255/UKDA-SN-9280-1.

64 See, for example: Leigh Shaw-Taylor and E.A. Wrigley, 'Occupational Structure and Population Change', in *The Cambridge Economic History of Modern Britain. Volume 1: 1700–1870*, ed. Roderick Floud, Jane Humphries and Paul A. Johnson (Cambridge University Press, 2014), 53–88; Leigh Shaw-Taylor, 'Diverse Experiences: The Geography of Adult Female Employment in England and the 1851 Census', in *Women's Work in Industrial England: Regional and Local Perspectives*, ed. Nigel Goose (Local Population Studies Society, 2007), 29–50; Kevin Schürer, Eilidh M. Garrett, Hannaliis Jaadla and Alice Reid, 'Household and Family Structure in England and Wales, 1851–1911', *Continuity and Change* 33, no. 3 (2018): 365–411; Kevin Schürer and Joe Day, 'Migration to London and the Development of the North–South Divide, 1851–1911', *Social History* 44, no. 1 (2019): 26–56.

65 See Olivia Vane's Observable notebook for Macromap, https://observablehq.com/@oliviafvane/macromap. This work formed the basis of this article: 'How Intrepid Victorian Surveyors Mapped the Length and Breadth of Britain', *The Economist*, 6 April 2023.

66 A.J. Christopher, 'The Quest for a Census of the British Empire c.1840–1940', *Journal of Historical Geography* 34, no. 2 (2008): 268–85; Alexander Kent, Soetkin Vervust, Imre Josef Demhardt and Nick Millea, *Mapping Empires: Colonial Cartographies of Land and Sea: 7th International Symposium of the ICA Commission on the History of Cartography, 2018* (Springer International Publishing, 2020).

67 A key example is the *Encyclopaedia of British Slave-ownership*, from the Centre for the Study of the Legacies of British Slavery, accessed 23 February 2026, https://www.ucl.ac.uk/lbs/project/details.

68 Evangeline Corcoran, Laura Siles, Smita Kurup and Sebastian Ahnert, 'Automated Extraction of Pod Phenotype Data from Micro-Computed Tomography', *Frontiers in Plant Science* 14 (2023): 1120182.

69 Blevins, 'Digital History's Perpetual Future Tense'. However, with this blithe statement we telescope the pipeline of work required to prepare these previously digitised materials for research. Despite two decades of investment, both public and private, in the digitisation of the content of our national libraries and archives, our experience has been that these assets are difficult to access, rarely research-ready, and surrounded by a mass of restrictions that have implications for reproducibility and the sharing of outputs. For a longer discussion, see Ahnert et al., *Collaborative Historical Research*, chapter 2.

70 On these initial research outcomes see Giorgia Tolfo, Olivia Vane, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, David Beavan and Katherine McDonough, 'The Minimum Research Outcome: A Mechanism for Generating and Managing Projects in Labs', in *Digital Humanities and Laboratories: Perspectives on Knowledge, Infrastructure and Culture*, ed. Urszula Pawlicka-Deger and Christopher Thomson (Routledge, 2023), 193–208.

71 See https://www.visionofbritain.org.uk/data/ and https://www.ordnancesurvey.co.uk/products/os-open-roads.

72 The code for this tool is available at https://github.com/Living-with-machines/MapReader. See also Kasra Hosseini, Daniel C.S. Wilson, Kaspar Beelen and Katherine McDonough, 'MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale', *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities* (2022): 8–19, https://doi.org/10.1145/3557919.3565812; and Kasra Hosseini, Katherine McDonough, Daniel van Strien, Olivia Vane and Daniel C.S. Wilson, 'Maps of a Nation? The Digitized Ordnance Survey for New Historical Research', *Journal of Victorian Culture* 26, no. 2 (2021): 284–299, https://doi.org/10.1093/jvcult/vcab009.

73 https://rchs.org.uk/railway-passenger-stations-in-great-britain-a-chronology/.

74 https://github.com/Living-with-machines/station-to-station.

Chapter 1

The Environmental Scan as a method for digital source criticism at scale

Kaspar Beelen, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Daniel van Strien and Daniel C.S. Wilson[1]

Introduction

When cultural artefacts are transformed into digital data, they often lose the context of their original creation. Until recently, many music streaming services stripped away contextualising information that had previously accompanied the physical record or CD. Classical music was particularly poorly served, even on platforms that offered lossless, hi-fi quality. Tracks might be labelled (and therefore discoverable) only by the performing artist or orchestra, rather than their original composer (e.g. J.S. Bach) or their title (e.g. Goldberg Variations).[2] And with tracks in any genre, information about who played what instruments, or about vocalists, session musicians, recording and release dates, or studio locations, is still rarely preserved, even though it was once easily to hand when the medium was a physical CD or record. Perhaps one day, the often-meandering original sleeve notes will be digitised in their entirety and made available as we listen. Providers may even go further and use AI to help structure the original information so that we can again see at a glance who sang that harmony or played that saxophone solo. Such loss may be little more than an inconvenience to the casual listener, but to a student of the history of music it represents a barrier that makes serious research impossible. Sadly, an equivalent loss of context bedevils almost all digitised collections.

Such contextualising information is known as 'metadata' and is critical for working with digitised collections of sources and documents. The loss of original metadata, which can situate documents in their proper context – in relation to their provenance as well as their social meaning – poses grave risks to users of digitised historical collections. As Lara Putnam has observed, while rapid digitisation has transformed access to historical sources via web access and keyword search, at the same time it has greatly diminished the importance of site-specific expertise and domain knowledge as a necessary precursor to making discoveries.[3] Just as the online listener risks being unable to know the who, when and where of their music stream, so also the scholar – whether using simple keyword searches or complex computational analysis – risks drawing the wrong, or at least radically decontextualised, conclusions from their queries.

Performing a keyword search is very different from manually browsing library shelves, catalogues and finding aids. Of course, both can be shaped by confirmation bias – we tend to find what we are looking for – but keyword search does not facilitate the chance discoveries that so often come when we flick through a physical item or browse the shelves around a volume. As a result, the sense we might have of the broader context of our findings is likely to be radically attenuated.

For example, Beelen recalls observing this phenomenon in action when, more than a decade ago, tutors switched to using the recently digitised Belgian Parliamentary proceedings for a political history seminar.[4] They noticed that, although searching was easier for the students, understanding the material was not. The processes required to find relevant information using analogue search functioned to equip students with crucial background knowledge, including source context, to help them critically interpret their findings. The students' analogue searches involved perusing reference materials, cross-referencing these with unwieldy printed indices. Though challenging for many, these practices provided a framework and context to make sense of the information eventually uncovered. By contrast, when students used a search box to access their digitised sources, they might quickly stumble upon multiple fragments relevant to their topic, but each document existed in a historical and archival vacuum.[5] Although information was easier to access, making sense of the results was more difficult. The difficulty was compounded by the scale of information (they often felt overwhelmed by search results), as well as its quality (poor optical character recognition (OCR) meant some

key sources had become invisible to their searches). This raised additional questions about what students did and did not find when navigating digitised collections.

Similarly, critical voices have begun to raise concerns about the distorting consequences for scholarly research, worrying that the ease of accessing digitised materials may inadvertently skew research priorities and influence academic arguments in unexpected ways.[6] There can be no simple, universal answer to this problem given the diversity of digital sources and the myriad research questions that can be addressed when using them. Different solutions will be required for different contexts, but there are some core principles that can help tackle the unintended distortions that can creep into digital methods if they are used uncritically. As well as adding important context to digitised collections, we set out some core principles for placing computational methods on a more robust footing by scaling up the critical methods central to the 'close reading' of sources so that they are also available to humanities scholars engaged in 'distant reading'.[7]

From the outset, the Living with Machines (LwM) team was determined to tackle these issues by addressing a fundamental question: how can humanities scholars' core techniques of source criticism be adapted for use at the much larger scale made possible by the digital revolution?[8] Faced with a report about, for instance, an election disturbance taken from a Victorian newspaper, the first thing a historian might ask is: what were the politics of this paper, where did it circulate and what readership was it aimed at? Often, the answers to such questions would be part of the historian's implicit knowledge, built up cumulatively from working intensively with a relatively small cluster of sources. But if the source was unfamiliar they would be able to consult a contemporary reference work to find out more. In the case of provincial newspapers, this would usually mean turning to one of the various annual '*press directories*' published from the late 1840s as a guide to advertisers and the publishing industry.[9] In LwM, we asked whether a similar approach could be applied at scale to an entire collection of digitised newspapers, both to improve digital searching and to better understand the collection's biases and representativeness. This was how the concept of the 'digital Environmental Scan' was born.[10] As the following narrative will show, while the principles at the heart of this methodology were developed first through our research on nineteenth-century newspapers, they are pertinent to any digital collection for which

good sources of supplementary reference metadata can be identified, digitised and processed.

To understand the need for such a digital Environmental Scan we must recognise that data, regardless of its scale or granularity, is always incomplete. Rob Kitchin describes data as 'oligoptic' rather than 'panoptic', challenging the notion – often prevalent in the sciences, but also sometimes in the humanities – that big data necessarily offers comprehensive and transparent views.[11] The massive increase in data collection during the 2000s led to bold claims of paradigm shifts and a new empiricism, the idea that instead of developing theories and conceptual frameworks, researchers need only find efficient ways to query big data for answers.[12] While such views have been widely criticised among humanities scholars and theorists, they have by no means disappeared. Researchers in computer science and AI sometimes need reminding that *bigger* data isn't necessarily *better* data, as issues of bias continue to surface in their models and applications.[13] This is what Kitchin means when he argues that 'all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible God's eye view'.[14]

Most historians would typically agree with this assessment: working at the individual document-level, they construct and refine their arguments by contextualising sources and determining what perspective these offer on the past. However, when faced with vast collections comprising millions of pages, or thousands of map sheets, historians lack the instruments to scale up their source criticism even though, when pressed, few would expect these resources to be neutral or impartial. Hence the key question is: how can we articulate this critical perspective when using extensive historical collections at scale, including interrogating what might be missing from our data?

This problem, it should be emphasised, is not particular to history but has wider application to many other areas of humanities research. The case studies presented in this chapter demonstrate how the Environmental Scan touches on diverse disciplinary fields including cartography and digital literary studies. The latter field, in particular, has long grappled with similar issues.[15] Katherine Bode has critiqued the ways distant reading and macroanalysis analyse literary works as abstract entities, ignoring insights into how these works gained meaning in specific contexts of production and reception.[16] Instead, digital corpora should capture the interconnections and model 'how literary works exist and relate to one another in

socially, spatially, and temporally relevant ways, not in terms of potentially, and certainly relatively, abstract categories of production'.[17]

The Environmental Scan represents a response to similar challenges of missingness, decontextualisation and abstraction. It began as a method to undertake the kind of source criticism at scale described above and was developed to work with the digitised newspaper collections, created through partnerships with the British Library over the last couple of decades (the Joint Information System Committee (JISC) digitised newspapers and the British Newspaper Archive (BNA)). This work developed into both a series of precise technical workflows for this particular use case and also a set of broader principles that could be applied to other sources and collections.

In the following pages we expound the broader values and principles that should be at the centre of any large-scale digital historical research. These values and principles are set out below, in the next section, through a discussion of our work on newspaper corpora. This section constitutes a complete argument, but, for those readers interested in other source types, they are invited to read two, shorter, following sections in which we extrapolate those principles to map collections (with examples drawn from our work on the nineteenth-century Ordnance Survey (OS) maps digitised by the National Library of Scotland (NLS)), and printed books (drawing on our work using the corpus digitised by Microsoft in partnership with the British Library). These additional examples demonstrate how the principles of the Environmental Scan can be generalised to very different collection types, if contextualising sources can be found.

In sharing an account of this research we have two aims. Firstly, we believe that the Environmental Scans that we undertook for these three collections represent important foundational work. This is why this chapter is placed at the beginning of this volume: it introduces the principles that grounded each of the sources that we touched, three of which are introduced as the case studies below. These sources are vital for the study of the nation's past, and therefore a rigorous understanding of the contours of these sources in their digital forms is of paramount importance. We hope that by drawing attention to the methods and datasets that we have generated people will deploy them in their own research. Secondly, by drawing out our method's underlying principles, we hope that people can adapt these for their own research. We see a variety of different potential audiences and beneficiaries. The kind of work we are advocating here

complements the existing work being done by GLAM (gallery, library, archive and museum) organisations like our project partners at the British Library, and may offer curators the ammunition they need for digitising reference works, releasing metadata or undertaking crowdsourcing projects that would facilitate the kind of analysis we discuss in the following pages. For other researchers already working in the space of computational humanities, or entering into it, we hope that they will begin to see this kind of source criticism as the precondition of rigorous analysis of the contents of digital collections. This is work that *must* be done before more focused research questions are tackled, for without a sense of the shape of our data, how can we draw sure conclusions?

In advocating for the necessity of source criticism at scale, we also hope to add our voice to the more pressing debate that is playing out around the development of AI technologies. While our examples all derive from digitised heritage collections, our arguments are equally pertinent to the training data fed into AI pipelines – and the stakes, of course, are much higher.[18] Indeed, we would contend that digital historical archives present a valuable testbed for illustrating the biases that are baked into any dataset, and the ways that we must seek to address those explicitly in the way that we design any research into, or uses of, such material. Here is an opportunity for the tech industry to learn from digital historians and curators.

Framing principles: the deep and wide scans

The term 'machines' in the 'Living with Machines' project's title carries an intentionally ambiguous meaning. It refers not only to the machines that emerged as part of the Industrial Revolution but also to how today's historians might engage with novel research instruments being developed by data scientists, including machine learning and other forms of artificial intelligence.

In this context, the Environmental Scan offers historians innovative tools for critically engaging with large heritage data. Regardless of size, data are always rooted in specific contexts and shaped by particular factors. However, when faced with vast amounts of digital data, these histories and contexts often become obscured, leaving researchers with a somewhat indecipherable agglomeration of information. When we query these datasets, such as through keyword searches, they will invariably yield some results – but often without providing context about the

source or, perhaps more importantly, about what might be missing. We cannot see what hasn't been digitised or know what portion of the collection is unsearchable because of mistranscription. This might sound self-evident, but it still needs to be kept in mind when working with digitised historical collections. The problems the digital Environmental Scan addresses are often acknowledged but rarely investigated, generally because historians (digital or otherwise) lack the tools and methods. The primary goal of the Environmental Scan is therefore to provide a framework and set of methods to contextualise large-scale digitised, historical data. It involves two key approaches which we might think of as the framing principles of the Environmental Scan: these are the 'deep scan' and the 'wide scan'.

The principle of the 'deep scan' concerns enrichment of the data, uncovering hidden or missing aspects of the data. Practically, this means enhancing or recovering information about the provenance of sources, including, where possible, their 'perspective'.

The deep scan acknowledges that while digital collections often include basic metadata, such as publication and/or digitisation date, they typically lack essential contextual information. For example, in digital newspaper collections, we may have access to the content but not to details about who originally created these materials, or important clues to who may have consumed them (e.g. price, declared politics, etc.). The deep scan seeks to fill this gap, and enrich digital collections, for example by associating the historical newspapers with metadata extracted from historical reference resources. This step thus offers richer descriptions and more detailed profiles of the collection. In their 2025 paper, 'Provocations from the Humanities for Generative AI Research', one of the provocations that Lauren Klein and her co-authors made was that:

> there is always a perspective that is shaped by training data, annotations, parameters, and prompts. Rather than chasing a 'representative' dataset – which, of course, can never be achieved – we would be better served by acknowledging that there are always perspectives encoded in our models, and by taking the time to document and name them.[19]

To summarise: the deep scan aims to enrich collections with additional layers of metadata, and thereby articulates the social, political and historical dimensions of digital objects that otherwise risk remaining opaque

and shapeless. Essentially, it aims to encode additional context informed by the user's research question.

The 'wide scan' situates digitised materials within a wider 'landscape' of information. In other words, it aims to gauge missingness. We use the metaphor of a 'landscape' to refer to the entirety of a particular source type – what scholars at Stanford Literary Lab have called 'the published' in contrast to 'the archive' and 'the corpus'.[20] In this case, all newspaper issues known to have existed within a specific spatial and temporal context. The wide scan estimates how the digital archive (or sample) relates to the population.

The wide scan encourages researchers to consider the composition and biases of digital collections, even when working with large-scale collections. It suggests that collections can be greatly enhanced not only by providing richer metadata (the deep scan) but also by providing detailed information on selection and missingness (the wide scan).

By applying the principles of the deep scan (data enrichment) and the wide scan (missingness and bias) to digitised collections, researchers can achieve the goals of large-scale source criticism. This approach transforms the way we critically examine digital archives and opens up new research possibilities in the humanities. It facilitates a more nuanced understanding of the limitations and potential of digital collections, allowing researchers to draw more accurate and contextually informed conclusions from their analyses. This involves not only creating new metadata fields but also describing 'absent' data. A significant part of our work relies on what we call 'historical reference metadata', which consists of data descriptions sourced from historical reference works. While the Environmental Scan involves processing, connecting and analysing data, a key aspect relies on ongoing discussions and collaborations with curators and archivists who manage reference works such as press directories and collection metadata.

Under the following headings we lay out the more precise steps and methods needed to implement these framing principles, and how these vary with different data types and specific collections. These methodologies enable a more robust and reliable examination of digital collections and are a crucial step toward making large-scale source criticism a reality.

Identify contextual sources

The digital Environmental Scan was developed initially through our work with the digitised newspapers. It allowed us to identify a more general

set of problems, as well as principles and workflow to address these. The first step in this was the identification of historical reference works which could be used to evaluate the composition and content of newspaper collections. These offer contextual information that is absent from current library systems.

In the case of the newspaper sources, our chosen reference source was the series of *Newspaper Press Directories*. The nineteenth century saw the emergence of a wide variety of directories, statistical compendia and other reference works seeking to offer systematic, scientific representations of society and economy to readers in business and the professions for whom knowledge was power.[21] As Laurel Brake notes, the *Newspaper Press Directories* were one of a range of such publications which emerged specifically to meet commercial needs, particularly the needs of publishers, advertisers and their agents.[22] The longest-running and most famous of these publications was Mitchell's *Newspaper Press Directory*, which began in 1846 and appeared annually from 1856. In 1856 the Post Office recognised Mitchell's *Directory* as an authoritative list of British newspapers. Although it faced competition from rival publications after 1870, Mitchell's retained its authoritative status well beyond the First World War.[23]

Newspaper directories played a crucial role in the history of media and advertising in the nineteenth century, collating newspapers' self-representations, codifying newspaper types, and facilitating the emergence of a knowable national press long before a few London-based titles came to dominate the newspaper landscape. As Tom O'Malley notes, from the outset directories sought to bring dignity and respectability to the newspaper industry: in its first edition Mitchell's declared its intention to offer 'a more dignified and permanent record' of the press than any previously available.[24] Mitchell's served the interests of advertisers, newspaper proprietors and journalists, providing detailed information on frequency of publication, area of circulation, pricing, political stance (if any), ownership and more. This is contemporary information not easily recovered from other sources, including standard library catalogues. It is this information that we have sought to extract and structure as enriched metadata.

Directories operated as a geographical index of the press, mainly distinguishing between the London and 'provincial' newspapers, which were in turn subdivided first by nation (the English, Scottish, Welsh and Irish press usually had separate sections), then by county and finally, at the lowest level, by town or district. As O'Malley notes, directories both described

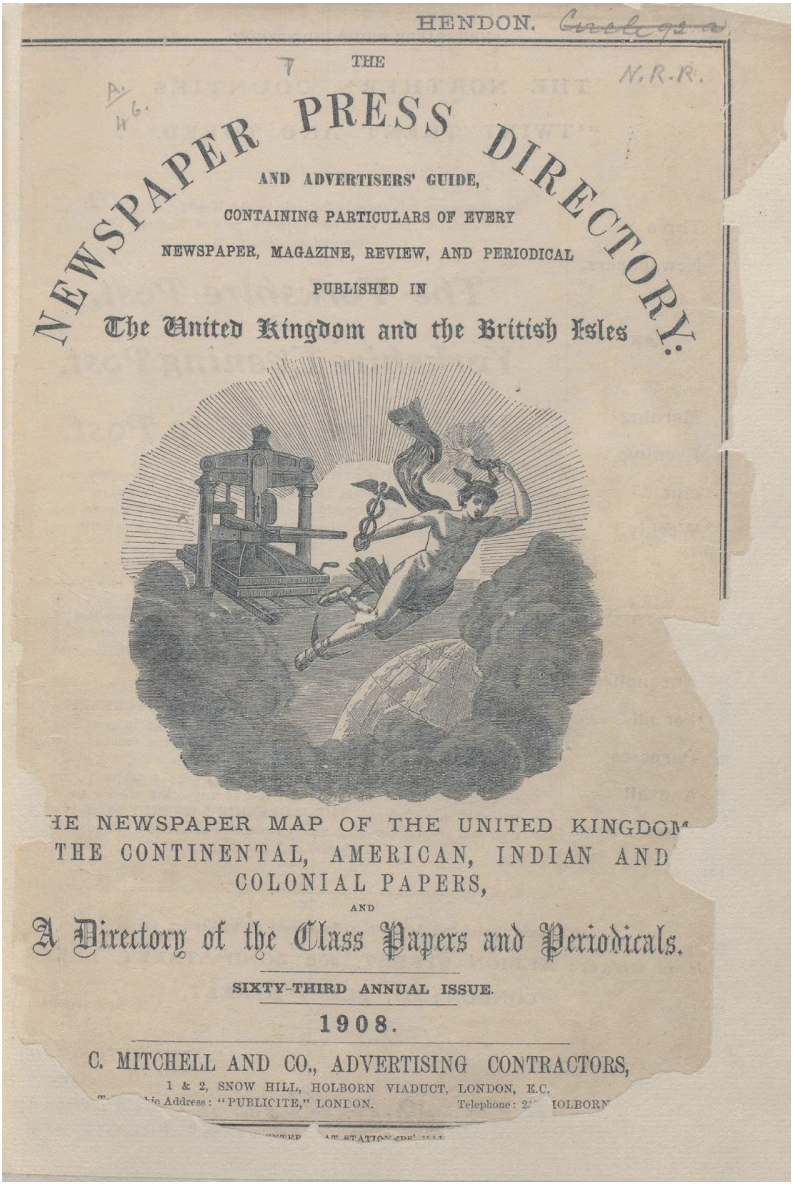
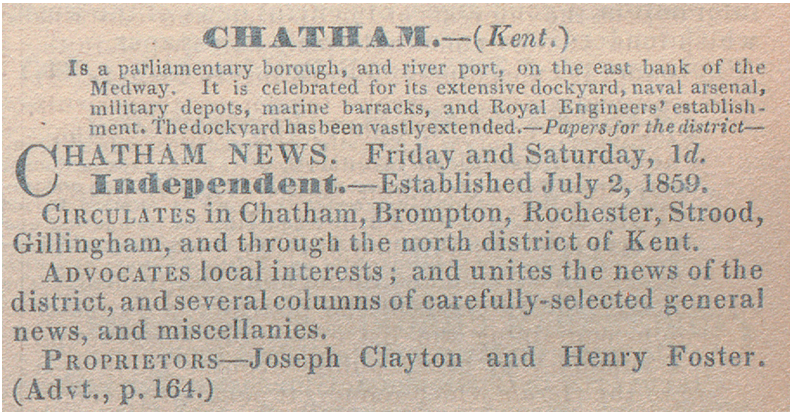


Figure 1.1a: Front page of the Mitchell's *Newspaper Press Directory* from 1908. Digitised by British Library Imaging Studio.



CHATHAM.—(*Kent.*)

Is a parliamentary borough, and river port, on the east bank of the Medway. It is celebrated for its extensive dockyard, naval arsenal, military depots, marine barracks, and Royal Engineers' establishment. The dockyard has been vastly extended.—*Papers for the district*—

CHATHAM NEWS. Friday and Saturday, 1*d*.
Independent.—Established July 2, 1859.
CIRCULATES in Chatham, Brompton, Rochester, Strood, Gillingham, and through the north district of Kent.
ADVOCATES local interests; and unites the news of the district, and several columns of carefully-selected general news, and miscellanies.
PROPRIETORS—Joseph Clayton and Henry Foster.
(Advt., p. 164.)

Figure 1.1b: An example of a newspaper description of the *Chatham News* in Chatham (Kent). Digitised by British Library Imaging Studio.[25]

and shaped the press with their typologies and their choice of features to foreground. Figure 1.1b shows an excerpt from Mitchell's 1908 *Directory*, describing an English title, the *Chatham News* from Kent. In addition to information about price, frequency, declared politics (here 'Independent'), longevity and ownership, each newspaper profile included information about where the paper could be purchased and the type of coverage it offered. Entries were often cross-referenced, as here, to fuller paid advertisements to be found within the *Directory*, and were preceded by brief descriptions of the locality. Our task on LwM was to find a way systematically to structure this rich information so that it could become 'enhanced metadata' for researchers seeking to work with digitised newspapers.

Create metadata from reference works

While it might be relatively simple to find directories or reference works to contextualise a dataset, turning these sources into metadata is non-trivial. Such works were designed to be read by researchers, not machines. The *Newspaper Press Directory* illustrates the complexity of the task: creating reference metadata from the bound volumes of the *Directories* involved multiple steps. In 2019, the LwM project requested the British Library Imaging Studio to digitise all editions of Mitchell's *Newspaper Press Directory*, spanning from 1846 to 1920. Led by Beelen, the project then further processed these scans to extract the text and make it machine-readable. The principal

goal was to convert scans of the pages into a structured format (i.e. tabular data). There were three separate steps to this:

- transform page images into text through OCR
- detect individual entries
- automatically parse the content of each entry for the attributes of the newspaper that would be tuned into enhanced metadata (i.e. information about newspaper titles, prices and political leanings, etc.).

As is often the case with heritage collections, this task proved to be more complex than initially expected, but we were keen to develop an automated method, rather than rely on manual transcription, in order to demonstrate the potential for applying this method to multiple collection types, regardless of size.

Although the typesetting followed a consistent pattern, Figure 1.2 illustrates the challenges the *Directories* posed for standard OCR software such as ABBYY in terms of unusual typeface layouts and the wide range of historical fonts deployed. Teams seeking to process analogous sources, especially when they are historical works, are likely to encounter similar problems. To overcome this, we fine-tuned off-the-shelf machine learning packages on test data until we had developed a model capable of automatically transcribing the *Directories*.[26] This method enabled us to develop a custom pipeline specifically tailored to the unique characteristics of the historical resource. In the first stage, we pre-processed the images to facilitate and improve text extraction. Beelen fine-tuned an image segmentation model (UNet) to split images into columns (so that the OCR would not read *across*

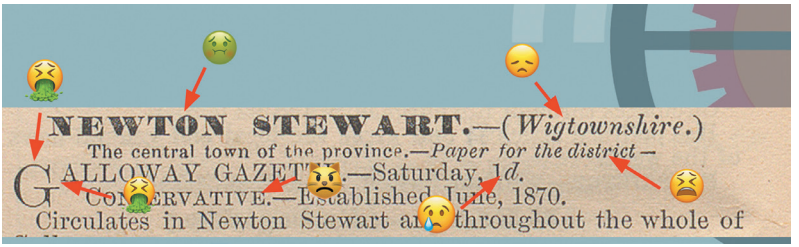
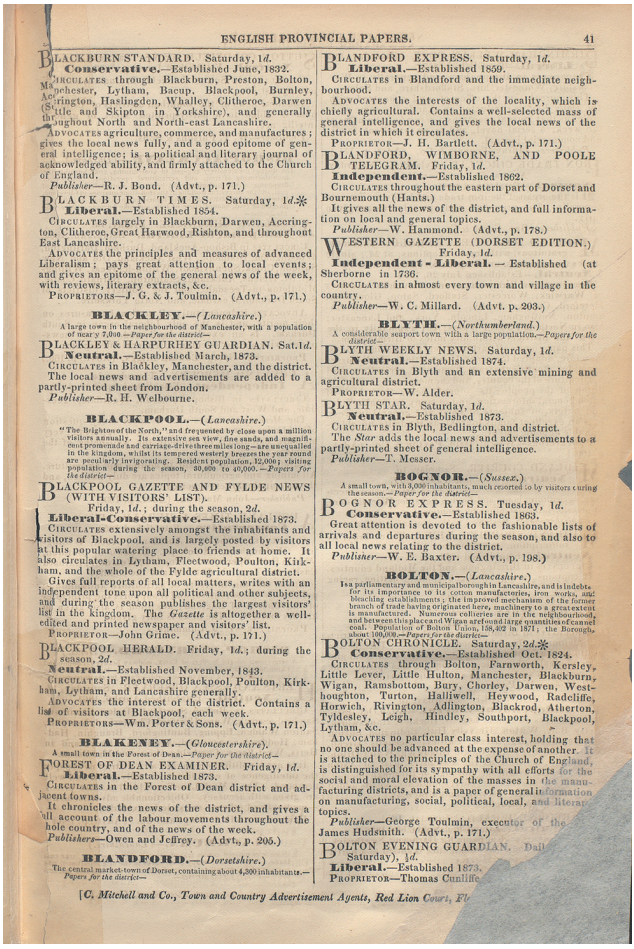


Figure 1.2: Issues with historical fonts and Optical Character Recognition. Digitised by British Library Imaging Studio.

columns) and address features that confused the OCR, such as the drop-down capitals at the start of each entry, which were 'shrunk' so that they could be lined up with the rest of the text.

Next, we fine-tuned an open-source OCR text recognition model (Tesseract 5) to address the issues with challenging fonts or unusual characters.[28]



ENGLISH PROVINCIAL PAPERS. 41

BLACKBURN STANDARD. Saturday, 1*d*.
Conservative.—Established June, 1832.
CIRCULATES through Blackburn, Preston, Bolton, Manchester, Lytham, Bacup, Blackpool, Burnley, Accrington, Haslingden, Whalley, Clitheroe, Darwen (Settle and Skipton in Yorkshire), and generally throughout North and North-east Lancashire.
ADVOCATES agriculture, commerce, and manufactures; gives the local news fully, and a good epitome of general intelligence; is a political and literary journal of acknowledged ability, and firmly attached to the Church of England.
Publisher—R. J. Bond. (Advt., p. 171.)

BLACKBURN TIMES. Saturday, 1*d*.*
Liberal.—Established 1854.
CIRCULATES largely in Blackburn, Darwen, Accrington, Clitheroe, Great Harwood, Risiton, and throughout East Lancashire.
ADVOCATES the principles and measures of advanced Liberalism; pays great attention to local events; and gives an epitome of the general news of the week, with reviews, literary extracts, &c.
PROPRIETORS—J. G. & J. Toulmin. (Advt., p. 171.)

BLACKLEY.—(*Lancashire.*)
A large town in the neighbourhood of Manchester, with a population of near 7,000.—*Paper for the district*—

BLACKLEY & HARPURHEY GUARDIAN. Sat. 1*d*.
Neutral.—Established March, 1873.
CIRCULATES in Blackley, Manchester, and the district. The local news and advertisements are added to a partly-printed sheet from London.
Publisher—R. H. Welbourne.

BLACKPOOL.—(*Lancashire.*)
"The Brighton of the North," and frequented by close upon a million visitors annually. Its extensive sea view, fine sands, and magnificent promenade and carriage-drive three miles long—are unequalled in the kingdom, whilst its tempered westerly breezes the year round are peculiarly invigorating. Resident population, 12,000; visiting population during the season, 30,000 to 40,000.—*Papers for the district*—

BLACKPOOL GAZETTE AND FYLDE NEWS (WITH VISITORS' LIST).
Friday, 1*d*.; during the season, 2*d*.
Liberal-Conservative.—Established 1873.
CIRCULATES extensively amongst the inhabitants and visitors of Blackpool, and is largely posted by visitors at this popular watering place to friends at home. It also circulates in Lytham, Fleetwood, Poulton, Kirkham, and the whole of the Fylde agricultural district.
Gives full reports of all local matters, writes with an independent tone upon all political and other subjects, and during the season publishes the largest visitors' list in the kingdom. The *Gazette* is altogether a well-edited and printed newspaper and visitors' list.
PROPRIETOR—John Grime. (Advt., p. 171.)

BLACKPOOL HERALD. Friday, 1*d*.; during the season, 2*d*.
Neutral.—Established November, 1843.
CIRCULATES in Fleetwood, Blackpool, Poulton, Kirkham, Lytham, and Lancashire generally.
ADVOCATES the interest of the district. Contains a list of visitors at Blackpool, each week.
PROPRIETORS—Wm. Porter & Sons. (Advt., p. 171.)

BLAKENEY.—(*Gloucestershire.*)
A small town in the Forest of Dean.—*Paper for the district*—

FOREST OF DEAN EXAMINER. Friday, 1*d*.
Liberal.—Established 1873.
CIRCULATES in the Forest of Dean district and adjacent towns.
It chronicles the news of the district, and gives a full account of the labour movements throughout the whole country, and of the news of the week.
Publishers—Owen and Jeffrey. (Advt., p. 205.)

BLANDFORD.—(*Dorsetshire.*)
The central market-town of Dorset, containing about 4,300 inhabitants.—*Papers for the district*—

BLANDFORD EXPRESS. Saturday, 1*d*.
Liberal.—Established 1859.
CIRCULATES in Blandford and the immediate neighbourhood.
ADVOCATES the interests of the locality, which is chiefly agricultural. Contains a well-selected mass of general intelligence, and gives the local news of the district in which it circulates.
PROPRIETOR—J. H. Bartlett. (Advt., p. 171.)

BLANDFORD, WIMBORNE, AND POOLE TELEGRAM. Friday, 1*d*.
Independent.—Established 1862.
CIRCULATES throughout the eastern part of Dorset and Bournemouth (Hants.)
It gives all the news of the district, and full information on local and general topics.
Publisher—W. Hammond. (Advt., p. 178.)

WESTERN GAZETTE (DORSET EDITION.)
Friday, 1*d*.
Independent - Liberal. — Established (at Sherborne in 1736.
CIRCULATES in almost every town and village in the country.
Publisher—W. C. Millard. (Advt. p. 203.)

BLYTH.—(*Northumberland.*)
A considerable seaport town with a large population.—*Papers for the district*—

BLYTH WEEKLY NEWS. Saturday, 1*d*.
Neutral.—Established 1874.
CIRCULATES in Blyth and an extensive mining and agricultural district.
PROPRIETOR—W. Alder.

BLYTH STAR. Saturday, 1*d*.
Neutral.—Established 1873.
CIRCULATES in Blyth, Bedlington, and district.
The *Star* adds the local news and advertisements to a partly-printed sheet of general intelligence.
Publisher—T. Messer.

BOGNOR.—(*Sussex.*)
A small town, with 5,000 inhabitants, much resorted to by visitors during the season.—*Paper for the district*—

BOGNOR EXPRESS. Tuesday, 1*d*.
Conservative.—Established 1863.
Great attention is devoted to the fashionable lists of arrivals and departures during the season, and also to all local news relating to the district.
Publisher—W. E. Baxter. (Advt., p. 198.)

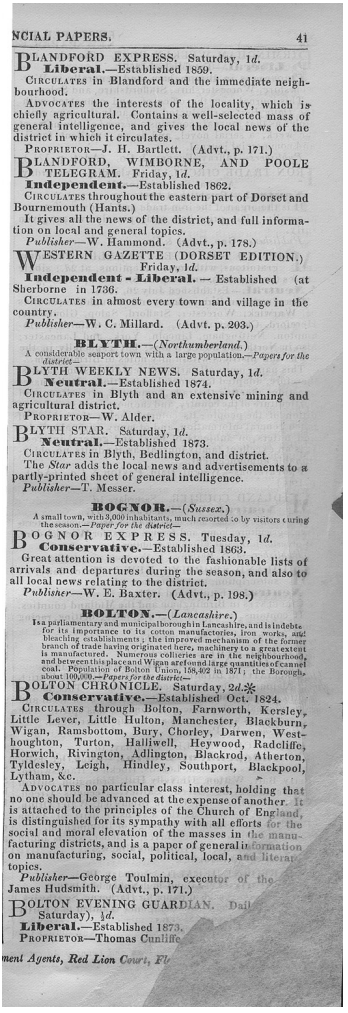
BOLTON.—(*Lancashire.*)
Is a parliamentary and municipal borough in Lancashire, and is indebted for its importance to its cotton manufactories, iron works, and bleaching establishments; the improved mechanism of the former branch of trade having originated here, machinery to a great extent is manufactured. Numerous collieries are in the neighbourhood, and between this place and Wigan are found large quantities of cannel coal. Population of Bolton Union, 158,402 in 1871; the Borough, about 100,000.—*Papers for the district*—

BOLTON CHRONICLE. Saturday, 2*d*.*
Conservative.—Established Oct. 1824.
CIRCULATES through Bolton, Farnworth, Kersley, Little Lever, Little Hulton, Manchester, Blackburn, Wigan, Ramsbottom, Bury, Chorley, Darwen, Westhoughton, Tarton, Halliwell, Heywood, Radcliffe, Horwich, Rivington, Adlington, Blackrod, Atherton, Tyldesley, Leigh, Hindley, Southport, Blackpool, Lytham, &c.
ADVOCATES no particular class interest, holding that no one should be advanced at the expense of another. It is attached to the principles of the Church of England, is distinguished for its sympathy with all efforts for the social and moral elevation of the masses in the manufacturing districts, and is a paper of general information on manufacturing, social, political, local, and literary topics.
Publisher—George Toulmin, executor of the late James Hudsmith. (Advt., p. 171.)

BOLTON EVENING GUARDIAN. Daily (Saturday), 1*d*.
Liberal.—Established 1873.
PROPRIETOR—Thomas Cunliffe.

[*C. Mitchell and Co., Town and Country Advertisement Agents, Red Lion Court, Fl*

Figures 1.3a, 1.3b and 1.3c: Image-processing pipeline: a) the original input page is shown, after which we extract columns and then shrink the drop-down capitals at the beginning of a newspaper description (b and c). Digitised by British Library Imaging Studio.[27]


NCIAL PAPERS. 41

BLANDFORD EXPRESS. Saturday, 1d.
Liberal.—Established 1859.
CIRCULATES in Blandford and the immediate neighbourhood.
ADVOCATES the interests of the locality, which is chiefly agricultural. Contains a well-selected mass of general intelligence, and gives the local news of the district in which it circulates.
PROPRIETOR—J. H. Bartlett. (Advt., p. 171.)

BLANDFORD, WIMBORNE, AND POOLE TELEGRAM. Friday, 1d.
Independent.—Established 1862.
CIRCULATES throughout the eastern part of Dorset and Bournemouth (Hants.)
It gives all the news of the district, and full information on local and general topics.
Publisher—W. Hammond. (Advt., p. 178.)

WESTERN GAZETTE (DORSET EDITION.) Friday, 1d.
Independent - Liberal. — Established (at Sherborne in 1736.
CIRCULATES in almost every town and village in the country.
Publisher—W. C. Millard. (Advt. p. 203.)

BLYTH.—(*Northumberland.*)
A considerable seaport town with a large population.—*Papers for the district*—

BLYTH WEEKLY NEWS. Saturday, 1d.
Neutral.—Established 1874.
CIRCULATES in Blyth and an extensive mining and agricultural district.
PROPRIETOR—W. Alder.

BLYTH STAR. Saturday, 1d.
Neutral.—Established 1873.
CIRCULATES in Blyth, Bedlington, and district.
The *Star* adds the local news and advertisements to a partly-printed sheet of general intelligence.
Publisher—T. Messer.

BOGNOR.—(*Sussex.*)
A small town, with 3,000 inhabitants, much resorted to by visitors during the season.—*Paper for the district*—

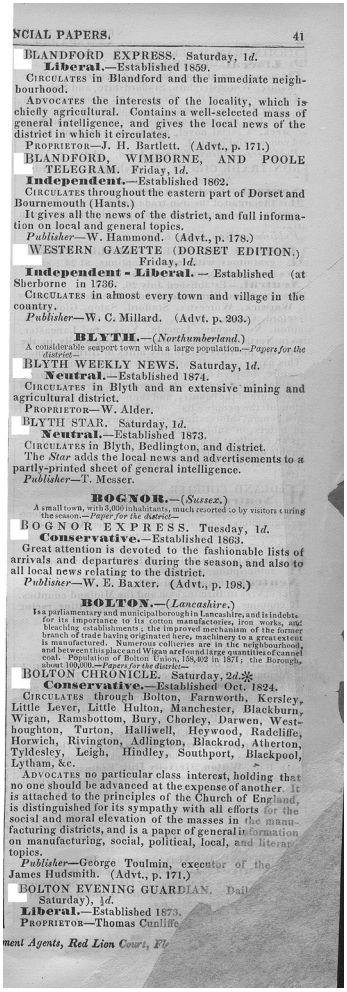
BOGNOR EXPRESS. Tuesday, 1d.
Conservative.—Established 1863.
Great attention is devoted to the fashionable lists of arrivals and departures during the season, and also to all local news relating to the district.
Publisher—W. E. Baxter. (Advt., p. 198.)

BOLTON.—(*Lancashire.*)
Is a parliamentary and municipal borough in Lancashire, and is indebted for its importance to its cotton manufactories, iron works, and bleaching establishments; the improved mechanism of the former branch of trade having originated here, machinery to a great extent is manufactured. Numerous collieries are in the neighbourhood, and between this place and Wigan are found large quantities of cannel coal. Population of Bolton Union, 158,472 in 1871; the Borough, about 100,000.—*Papers for the district*—

BOLTON CHRONICLE. Saturday, 2d.*
Conservative.—Established Oct. 1824.
CIRCULATES through Bolton, Farnworth, Kersley, Little Lever, Little Hulton, Manchester, Blackburn, Wigan, Ramsbottom, Bury, Chorley, Darwen, Westhoughton, Turton, Halliwell, Heywood, Radcliffe, Horwich, Rivington, Adlington, Blackrod, Atherton, Tyldesley, Leigh, Hindley, Southport, Blackpool, Lytham, &c.
ADVOCATES no particular class interest, holding that no one should be advanced at the expense of another. It is attached to the principles of the Church of England, is distinguished for its sympathy with all efforts for the social and moral elevation of the masses in the manufacturing districts, and is a paper of general information on manufacturing, social, political, local, and literary topics.
Publisher—George Toulmin, executor of the James Hudsmith. (Advt., p. 171.)

BOLTON EVENING GUARDIAN. Daily Saturday), 1d.
Liberal.—Established 1873.
PROPRIETOR—Thomas Cunliffe.

ment Agents, Red Lion Court, Fl


NCIAL PAPERS. 41

BLANDFORD EXPRESS. Saturday, 1d.
Liberal.—Established 1859.
CIRCULATES in Blandford and the immediate neighbourhood.
ADVOCATES the interests of the locality, which is chiefly agricultural. Contains a well-selected mass of general intelligence, and gives the local news of the district in which it circulates.
PROPRIETOR—J. H. Bartlett. (Advt., p. 171.)

BLANDFORD, WIMBORNE, AND POOLE TELEGRAM. Friday, 1d.
Independent.—Established 1862.
CIRCULATES throughout the eastern part of Dorset and Bournemouth (Hants.)
It gives all the news of the district, and full information on local and general topics.
Publisher—W. Hammond. (Advt., p. 178.)

WESTERN GAZETTE (DORSET EDITION.) Friday, 1d.
Independent - Liberal. — Established (at Sherborne in 1736.
CIRCULATES in almost every town and village in the country.
Publisher—W. C. Millard. (Advt. p. 203.)

BLYTH.—(*Northumberland.*)
A considerable seaport town with a large population.—*Papers for the district*—

BLYTH WEEKLY NEWS. Saturday, 1d.
Neutral.—Established 1874.
CIRCULATES in Blyth and an extensive mining and agricultural district.
PROPRIETOR—W. Alder.

BLYTH STAR. Saturday, 1d.
Neutral.—Established 1873.
CIRCULATES in Blyth, Bedlington, and district.
The *Star* adds the local news and advertisements to a partly-printed sheet of general intelligence.
Publisher—T. Messer.

BOGNOR.—(*Sussex.*)
A small town, with 3,000 inhabitants, much resorted to by visitors during the season.—*Paper for the district*—

BOGNOR EXPRESS. Tuesday, 1d.
Conservative.—Established 1863.
Great attention is devoted to the fashionable lists of arrivals and departures during the season, and also to all local news relating to the district.
Publisher—W. E. Baxter. (Advt., p. 198.)

BOLTON.—(*Lancashire.*)
Is a parliamentary and municipal borough in Lancashire, and is indebted for its importance to its cotton manufactories, iron works, and bleaching establishments; the improved mechanism of the former branch of trade having originated here, machinery to a great extent is manufactured. Numerous collieries are in the neighbourhood, and between this place and Wigan are found large quantities of cannel coal. Population of Bolton Union, 158,472 in 1871; the Borough, about 100,000.—*Papers for the district*—

BOLTON CHRONICLE. Saturday, 2d.*
Conservative.—Established Oct. 1824.
CIRCULATES through Bolton, Farnworth, Kersley, Little Lever, Little Hulton, Manchester, Blackburn, Wigan, Ramsbottom, Bury, Chorley, Darwen, Westhoughton, Turton, Halliwell, Heywood, Radcliffe, Horwich, Rivington, Adlington, Blackrod, Atherton, Tyldesley, Leigh, Hindley, Southport, Blackpool, Lytham, &c.
ADVOCATES no particular class interest, holding that no one should be advanced at the expense of another. It is attached to the principles of the Church of England, is distinguished for its sympathy with all efforts for the social and moral elevation of the masses in the manufacturing districts, and is a paper of general information on manufacturing, social, political, local, and literary topics.
Publisher—George Toulmin, executor of the James Hudsmith. (Advt., p. 171.)

BOLTON EVENING GUARDIAN. Daily Saturday), 1d.
Liberal.—Established 1873.
PROPRIETOR—Thomas Cunliffe.

ment Agents, Red Lion Court, Fl

Figures 1.3b and 1.3c: (cont.).

The idea was to train the model on those tricky cases with manual transcription so that it could 'learn' how to deal with those cases. We then applied the fine-tuned Tesseract to previously segmented and processed images (see the first step in the previous list). Overall, we found that combining image segmentation with fine-tuned OCR markedly enhanced the quality of the text extraction.[29]

Finally the OCR transcriptions were processed in two stages using text segmentation and semantic annotation to extract the attributes we are interested in as enriched metadata for each newspaper title (e.g. political leaning, price, ownership, self-declared interests, etc.). First, we segmented the OCR text to divide the pages into individual newspaper entries. From these descriptions we extracted specific elements such as title, price and political leaning. For both steps we fine-tuned BERT language models on manually annotated data. The final output was converted into a tabular format, with each row representing a newspaper profile and each column recording that newspaper's specific attributes.[30] Subsequently, we post-processed the output by normalising values for fields such as title, political leaning, price and place of publication (the latter, as we will see in Chapter 5, forms a vital field for examining how change in the use of certain technological words proceeded differently in the north and south of England).[31] Finally, Beelen manually corrected the data, paying particular attention to titles and locations that were difficult for the OCR engine to handle, and supervised the record linkage as needed.

Link contextual metadata

The final stage of data preparation that is required before the digital Environmental Scan can be undertaken is data linkage. The kinds of linking required may be threefold, depending on the nature of the source: the first is linking of entries within the contextual dataset; the second is the linkage of the entities within the contextual source with the digitised collection under investigation; and the third is the addition of Linked Open data to facilitate the further interoperability of the data.

The rich metadata extracted from each issue of the *Newspaper Press Directories* provides a rich account of the newspaper landscape for each year (1846–1920). The first task was to link the records for each newspaper title over time. For other data types, this may not be an issue, but for any sources that were released as part of a series, temporal linking will be a required step. For this purpose, we developed a record linker by training a BERT language model on annotated title pairs, which we labelled as 'same' or 'different'. To link an entry e in year t to an entry in year $t + 1$ we fetch ten similar titles to e in year $t + 1$ and then classify all pairs of $(e, e_t + 1_i)$ with $(1 < i < 10)$ and take the one with the highest probability for 'same' (if any are classified as 'same').[32] Finally, Beelen manually corrected the data, paying particular attention to titles that were difficult for the OCR

engine to handle, and supervised the record linkage as needed.[33] Linking resulted in most of the entries being assigned a 'chain identifier' which groups records as describing the same title over time.

After internal linking (i.e. grouping entries within the collection) we also proceeded to associate *Directories* with external data (i.e. the digitised newspapers themselves). This was done semi-manually. For each digitised newspaper, we tried to link it to the most appropriate 'chain identifier' in Mitchell's based on the similarity of the title, place of publication and established data. We ranked the ten most likely matches found in Mitchell and picked the best candidate. For newspapers published before 1846, we used backward linking by attempting to match digitised newspapers to the earlier editions of the *Newspaper Press Directories*.

There is also the opportunity to link the data to other knowledge bases. In our case we were able to link placenames to their Wikidata IDs and geocoordinates. We did this using the toponym linking tool T-Res, that we developed on LwM.[34] Given an input text, T-Res identifies the places that are mentioned in it ('The Recogniser' component of the pipeline), identify candidate matches, and ranks them by probability ('The Ranker'), links them to their corresponding Wikidata IDs, and provides their geographic coordinates ('The Linker'). The benefit of adding Linked Open Data like this is that it opens up the dataset to richer modes of analysis. In this case it enables us to explore the spatial distribution of our data, and other geospatial patterns that may be latent in our data.

This brief overview shows the considerable labour that was entailed first to turn the text of the *Newspaper Press Directories* into tabular data, rationalising the data through temporal linking, enriching it and finally linking it to the digital collections under analysis. To summarise, the digitised *Directories* data contains 92,000 descriptions of approximately 6,600 newspaper titles published between 1846 and 1920. For each title, we systematically extracted information on politics, price, location and people, among other aspects. Place names associated with publication and circulation were georeferenced, and entries were linked across time to facilitate the study of a newspaper's evolution. It perhaps explains why more of this work has not been done either by research projects or GLAM institutions. However, the resulting dataset is substantive and ripe with potential, and has already enabled vital new insights which we discuss in the 'Map the Landscape' section below, as well as in Chapter 5, 'Analysing the Language of Mechanisation in Nineteenth-Century British

Newspapers'. Importantly, all this data has been made publicly available, with a sample shown in Table 1.1, with the intention of enabling others to make the Environmental Scan a foundational step in their own research on nineteenth-century newspapers.[35] But what do the foundational principles of the Environmental Scan look like in action?

Map the landscape

As the previous account shows, creating reference metadata from historical sources can be a time-consuming and complex process. But the opportunities they unlock are substantial. We suggest in this section that, before even turning to the process of contextualising our digitised collections, the first step of the Environmental Scan should be to use the new reference dataset to map the landscape of historical data as documented in these sources. If the process of archiving the past was perfect, and nothing listed in these reference works had been lost, what would the data landscape have looked like? We found that analysing the data from the *Newspaper Press Directories* alone already gave us some powerful insights into the evolving information landscapes of the past.

As listed above, we collected many attributes about the different newspaper titles, from price, to topics of interests, and of course self-declared political leaning. Focusing on just this last attribute listed in Mitchell's *Newspaper Press Directory* illuminates the political dynamics of the Victorian press and how these changed over time. The majority of Victorian newspapers registered a partisan affiliation, with approximately half identifying as either 'liberal' or 'conservative', thereby associating themselves with the politics of the two main political parties of the era, Liberal and Conservative, although it should be acknowledged that the meaning of 'party' was itself in flux between the 1830s and 1880s.[36] Reflecting this fluidity, large numbers chose other partisan labels such as 'liberal conservative' or 'independent liberal', and many (almost 40 per cent) claimed to be either 'neutral' or 'independent' (not simple synonyms in this typology: independent newspapers offered much greater coverage of politics than 'neutral' newspapers). The extensive and detailed character of Mitchell's political typology points to the role that newspapers played in shaping how they were listed in the *Directories*. Although the principal labels of 'liberal', 'conservative', 'neutral' and 'independent' dominate, the *Directories* contain a long tail of labels, allowing for finer nuances of political position to shine through (Figure 1.4). In total, we recorded seventy different political labels.[37]

Table 1.1: Selected examples from the structured *Newspaper Press Directories* showing the extracted data for the *Brackley Observer* and *Altrincham and Bowdon Guardian*

Title	Politics	District	County	Price	Date	Description	Link_id
Altrincham and Bowdon Guardian	neutral	Bowdon	Cheshire	2d; 3d	1862	Takes no part in politics. [. . .].	null
Brackley Observer	liberal	Brackley	Buckinghamshire	1½d	1858	Taking a liberal view of politics . [. . .]	CID_00570

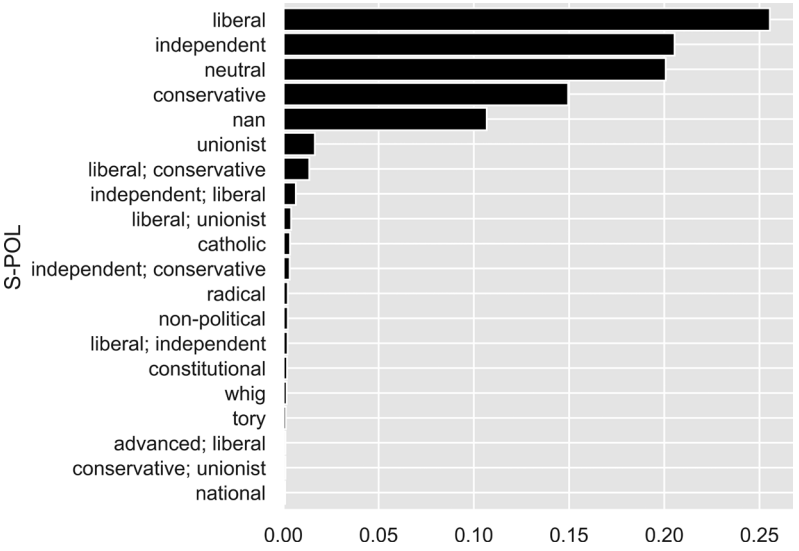


Figure 1.4: Distribution of political labels. Credit: Kaspar Beelen.[38]

But what do these labels signify? Are they persistent identifiers of a newspaper's political orientation? Or should we treat them more cautiously, as categories that are permeable? To answer these questions, we calculated the frequency of label transitions (i.e. of newspapers changing their political affiliation between editions of the *Directories*). As explained in the previous section, we linked the annual records of newspaper titles over time, which allowed us to study their diachronic evolution. Based on this, we computed the extent to which newspapers changed their orientation over time. Table 1.2 shows the ten most frequent transitions.

Inspecting the changes in political affiliation over time allowed us to investigate the use of these labels. We observed that political orientation is not a permanent fixture: one-quarter of the newspapers that appear more than once changed their orientation during their run (the proportion was even higher for newspapers that reappear more than three times, running to 35 per cent). Table 1.2 lists the most frequent transitions, indicating that even though titles rarely 'crossed the floor' – in the sense of switching directly from liberal to conservative or vice versa (only forty-five in total) – they did adapt their political orientation over time.[39] The neutral and independent categories appeared to be more fluid and these labels were

Table 1.2: Frequency of the ten most common shifts in political orientation

Source (was)	Target (became)	Count
Neutral	Liberal	127
Neutral	Independent	113
Independent	Liberal	108
Liberal	Independent	105
Neutral	Conservative	67
Liberal & Conservative	Conservative	43
Conservative	Independent	40
Conservative	Liberal & Conservative	29
Liberal	Conservative	29
Independent	Unionist	26

most susceptible to change. Not only were they more interchangeable among themselves (we noted 133 switches in total) but also with liberal and conservative titles.

A temporal reading of this typology suggested that the political orientation of newspapers was far from set in stone, but some borders were more porous than others. The somewhat elusive category of independent and neutral periodicals emerged as playing a critical role. Not only did they absorb large swathes of the more partisan press but they were also the source from which many liberal or conservative papers later emerged. Put in numbers: we noted 229 transitions of liberal/conservative to neutral/independent, and 375 in the other direction. In other scenarios, the changes remain relatively modest but nonetheless revealing. For example, many transitions between 'liberal & conservative' to conservative (and vice versa) suggest that 'liberal' was added as a qualifier to the type of conservatism a newspaper advocated, rather than as a way to signify support for both parties (this is consistent with what we know about divisions in mid-Victorian conservative politics).[40] The conservative press seemed more resilient compared to its liberal counterpart, with the latter appearing almost twice as often as the source of a transition. The great Liberal split over Irish Home Rule in 1886 was only part of this story (although it

doubtless also explains the switch of twenty-six 'independent' newspapers to 'unionist' in Table 1.2).

What drives these transitions? For now, we can only speculate on the reasons. Did institutional factors, such as changes in ownerships or news editors, play a crucial role? Did politics of the press shift in line with regional or national electoral trends? By releasing the digitised and structured *Newspaper Press Directories*, we hope future researchers can both improve the data and harness it to shed new light on the historical evolution of the British press.

Thirdly, we investigated the relationship between newspapers' political labels and the longer description of each newspaper published in the *Directories*. We estimated the association between metadata (here political label) and topic prevalence using a structural topic model, which allowed us to identify distinctive topic clusterings in the descriptions of each newspaper.[41]

By incorporating metadata in the model, the structural topic model enabled us to infer the strength of association between topics and document-level variables. Figure 1.5 plots the change in topic proportions shifting from liberal (left) to conservative (right). Only, two topics appeared to be overtly political. The topic cluster '"interest", "agricultur", and "advoc"' tilted quite strongly towards conservatism, reflecting the roots of that party in the landed and agricultural interest, and the topic cluster '"advoc", "polit", and "liber"' tilted to liberal, but less strongly so, perhaps indicating that liberalism was more invested in politics as a process and more overtly political.[42]

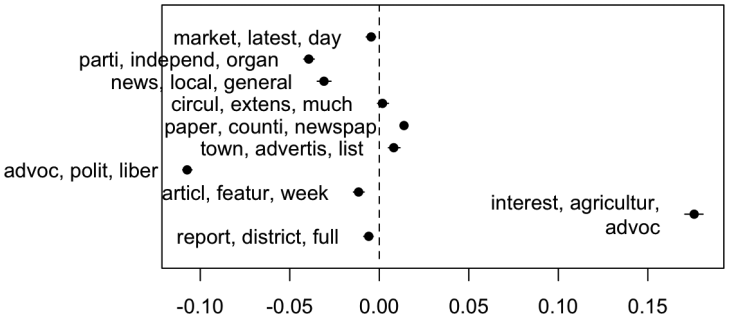


Figure 1.5: Topics positioned by the strength of their association with liberal (left) or conservative (right). Credit: Kaspar Beelen.

Most of the topics in newspaper descriptions were centred around zero, meaning that they were largely neutral with respect to the political orientation of the newspapers. The effect size for other topics (e.g. '"parti", "independ", and "organ"') was statistically significant but still relatively small. Arguably, local newspapers felt commercially obliged to signal their core partisan allegiance (to agricultural interests or political reform respectively) but also wished to signal that the bulk of the content was of general interest to *all* readers. Put differently: both conservative and liberal newspapers foregrounded many of the same features of their coverage, such as the scope and circulation of the paper, the local interests it promoted or other characteristics (such as good connections by rail or telegraph) instead of political features. Interestingly, we found that the tendency to downplay the newspapers' distinctive political features in these portraits increased over time. Even though Mitchell's continued to list newspapers' political orientation as a key attribute, from a discursive point of view, the portraits of liberal and conservative newspapers became less distinguishable from the 1880s, suggesting that, for most newspapers, the desire to maximise readership increasingly triumphed over partisan considerations.[43] Indeed, as Nicolò Bonato's work on the dataset illustrates, by 1922 local newspapers across much of Britain had dropped partisan allegiances, preferring instead to present themselves as 'independent'.[44]

The aim here is not to fully narrate the insights and findings from this analysis of the self-declared political leanings of the respective newspapers. Rather, the point is to illustrate how historians can reconstruct the landscape of information about the past using reference works. Our analysis uses elements of the 'deep' scan (enhanced metadata derived from contemporary reference sources) to create the potential for conducting a 'wide' scan. It does this by constructing a richer picture of the known press *as a whole* (rather than only digitised titles), what we have called the press landscape. Using the historical classifications and language extracted from contemporaneous reference works makes it possible to conduct time-sensitive, contextualising 'wide' scans of digitised collections (i.e. comparing these collections to the shifting landscape of all known newspapers at any particular moment). Not only does this analysis provide important insights into the use and meaning of the labels and categorisation practices that underlie the enriched metadata but it also sketches wider changes in the press landscape that can inform the interpretation of newspaper content.

This brief preview of our findings shows what is possible from just a single attribute extracted from the *Newspaper Press Directories*: self-defined partisanship. Through the lens of just this one variable we gauge how systems of classification and political alignments changed over time. But newspaper directories recorded many other attributes that can be leveraged to deepen our understanding of both the historic newspaper landscape and the characteristics of digitised newspaper collections. Perhaps the most important extra variable is price point. This can be used to distinguish between newspapers aimed at a mass market (typically priced at a penny (1d), or even half a penny, by the late nineteenth century), and those costing more and targeted at a narrower but wealthier readership. Given the absence of reliable information on either circulation or readership demographics, this variable makes it possible to conduct socio-economic analyses of historic newspapers that would otherwise prove impossible.

Measure the contours and biases of the digital landscape

The foregoing steps and principles are the preconditions of the task which we set out in this chapter: to enable a *digital* Environmental Scan, to understand the representativeness of the digital collection with which we now work in relation to the landscape of potential historical data. In our case, we can use the historical metadata collected from Mitchell's *Newspaper Press Directory* not only to reconstruct the broad contours of the newspaper landscape of the past but also to analyse the representativeness, and potential hidden biases, of major British digitised newspaper collections.[45] Such an approach should combine elements of the 'deep' and 'wide' scans. In the case of the digitised newspaper collections that might mean leveraging enhanced metadata attributes (the 'deep scan'), while also contextualising our corpus of digitised newspapers in relation to the overarching contemporary newspaper landscape (the 'wide scan').

Our first experiment in applying this principle was on the JISC corpus of digitised provincial newspapers. This corpus was funded by the UK's JISC in a collaboration with the British Library and ran between 2004 and 2009.[46] The collection now forms part of the much larger BNA. We chose to work with it mainly because of its small size, which was ideal for testing purposes, and because it is readily available to researchers via British Library Labs, but we were also aware that it had been specifically designed to meet the perceived needs of academic users.[47] Conducting a wide Environmental

Scan would allow us to delineate how these needs had shaped the JISC collection in comparison to the broader press landscape.

In short, we were interested in recording the representativeness and bias of the JISC digital collection. 'Bias' is of course a loaded term and often used pejoratively, but we are employing it here in the statistical sense of a systematic distortion. We computed this statistical bias using measures derived from information theory by calculating the divergence (specifically the Jensen-Shannon Divergence) in label distributions between the wider 'landscape' of newspapers reconstructed from Mitchell's and the digital collection. Focusing on the political labels, we wanted to see if the same proportions of labels were present in the landscape and the digital collection, and whether these distributions tracked one another over time.[48] Using these measures of bias, we observed that the JISC digital collection exhibited a conservative bias, and under-represented neutral and independent newspapers, leading to an increasing divergence between the sample and the actual newspaper landscape over time. Specifically, the JISC corpus under-sampled cheaper newspapers aimed at a mass readership, especially during the rapid expansion of the popular press in the 1860s and 1870s. In general, it over-represented politically aligned newspapers at the expense of neutral and independent titles, potentially exaggerating the politicisation of the Victorian public sphere. Moreover, while the JISC corpus offered broad nationwide coverage, it under-represented the south-east of England and some large counties.

We would contend that the urgency of such an approach increases with the size of the data, in part because people are easily lulled into a false sense of security when data is 'big'. One might imagine that since JISC was an older, relatively small-scale digitisation project some of its biases and distributions may have been redressed by the waves of newspaper digestion that have happened since the 2000s. For this reason, we turned to the BNA.[49] The BNA is an ongoing effort in partnership between the British Library and FindMyPast to digitise the BL's newspaper holdings. In 2019 FindMyPast shared with the LwM project a copy of the digitised collection at that date containing 1,205 titles. The complete newspaper dataset (including JISC, Heritage Made Digital and further titles digitised for the project and subsequently added to the BNA) comprised 1,400 individual titles, and approximately 150 billion words for the period 1790–1920.[50]

The digital Environmental Scan allowed us to observe that the BNA digital collection still over-represented partisan newspapers, but

this over-representation of liberal and conservative titles compared with the wider press landscape recorded in Mitchell's *Directory* has been considerably reduced when compared with the original JISC collection. This probably reflects a shift in the primary market for digitised newspapers from academic researchers, especially historians, to family historians and genealogists. JISC's purpose in digitising newspapers had been to identify long-running, influential titles that came closest to being regional 'newspapers of record'. But based on our analysis, priorities slowly shifted when FindMyPast took over development of the collection. Not only did the rate of digitisation increase markedly but the selection also changed, judging by the much larger numbers of neutral and independent titles that were digitised. Often viewed as more parochial, and even trivial, by academic historians, such titles arguably represent a better source for those looking for traces of their ancestors in the archive. Perhaps unintentionally, this shift from influential, often party-aligned titles to more parochial ones means that the digital corpus is coming more closely to reflect the tastes of Victorian readers, who as Andrew Hobbs argues, generally placed a strong premium on local interest over national news, including national politics.[51]

But what might we do with this information? In working with our colleagues in GLAM organisations, we would contend that this information is useful for helping organisations to think more holistically about their digitisation priorities going forwards. Of course, redressing all the imbalances documented by an Environmental Scan is in many ways an impossible task. Like Lauren Klein and her co-authors, we believe that it is more important to recognise that 'there are always perspectives encoded in their data and models', and our task should be to 'take time to document and describe them'.[52] This is what the digital Environmental Scan makes possible. Like Klein et al., rather than advocating for the elimination of bias in digital collections, we suggest a way to articulate those biases through enhanced metadata that researchers can then use to guide their research questions. In an ideal world, these kinds of contextual data might be baked into future digital collection releases by GLAM organisations and their partners, allowing for greater transparency about their data. Such data could also guide a more self-aware process of digitisation: if the aim is to build digital collections that echo the contours of the historical data landscape, then aligning proportionately against key attributes of the data might be one way to achieve this. Collections could be made more representative even if a completely representative sample is unattainable. Organisations could also use the

data to curate digital subsamples of materials that foregrounded specific audiences, for instance newspapers supporting Irish nationalism, or claiming to be written for female readers or the working classes. Deciding *which* attributes to prioritise would be a policy decision for curators in consultation with the primary users of these resources.

Reframe analyses

The Environmental Scan is not just a tool for describing the broad contours of digital collections resulting from digitisation policies. Using the enhanced metadata, this framework also enables researchers to contextualise content analysis of their sources. As we'll see in Chapter 5, which analyses the language of mechanisation in nineteenth-century British newspapers, there appear to be differences in the semantic trajectories of words over time – such as 'bulb', 'coach', 'gear', 'matches', 'railway', 'stamp', 'trade', 'traffic', 'train' and 'wheel' – between north and south England. The metadata also allows us to understand how the provenance of the dataset may shape the outcomes of research queries.

For example, we undertook content analysis of the *Newspaper Word Count Corpus* – a dataset we developed in the project, which comprises a gigantic collection of word counts associated with enriched metadata.[53] This allowed us to analyse word usage in its historical context. Based on these data we could address questions such as: does the word 'worker' appear more in the liberal than the conservative press? Where and when does the word 'drought' appear? We constructed the corpus by collecting monthly word counts for each title, and linking these counts to the *Newspaper Press Directories*, thereby adding context to these counts.

Studying text in relation to our refined and historical metadata allowed us to perform more complex analysis of newspaper content. We were especially interested in the political biases, that is, figuring out how language varied depending on the self-declared political affiliation of newspapers. Using an adapted version of the Fightin' Words algorithm, we computed which words were most 'partisan', that is, indicative of newspapers' declared political allegiances.[54] The Environmental Scan enabled us to foreground the partisan nature of newspaper language. We saw that 'tory' and 'liberal' emerged as the most distinctive words used by newspapers self-declaring as liberal, whereas 'radical' and 'conservative' were most distinctive for the conservative press. Drawing discursive borders between 'we' and 'they', between in- and out-groups, was clearly part of newspaper

language adding further weight to our earlier findings, namely that the digitised collection, which already appeared to oversample political newspapers, might lead to an overstated prevalence of partisan viewpoints.

By contextualising our text analysis within the enriched metadata, we were also able to observe another important skew on the dataset: there was an observable correlation between OCR quality, political orientation and price. Poor OCR is not evenly distributed across the digital corpus. Our lists of typical conservative partisan words tended to contain very few OCR errors when compared to other types of newspaper, especially the neutral and liberal ones. At the same time, lists generated from more expensive newspapers carried fewer errors than those from cheap newspapers. It seems likely that these findings reflect the poorer quality paper and ink used for cheaper newspapers, rather than anything intrinsic to the digitisation process. However, differential OCR quality matters once it's not random. Here it further accentuates biases we had already identified in digital collections, suggesting that researchers could be unwittingly privileging specific partisan perspectives, namely elite and Conservative, regardless of whether they are conducting a simple keyword search or computational linguistic analysis. As we suggest above, when historians are working on the level of single documents or issues of a newspaper, this is the kind of contextualisation they can easily do. The Environmental Scan puts that kind of contextualisation back in reach for researchers working at the most distant level of reading, that is, on whole collections of digitised newspapers.

Use enhanced metadata to model alternate worldviews

This brief account of our research on digitised newspaper collections suggests how the Environmental Scan can help researchers think about how bias should be measured; how we understand the hidden amplification and some perspectives and the attenuation of others. However, it is vital that this is framed within a more capacious and informed assessment of what 'bias' and 'representativeness' mean. As explained above, our initial bias measures focused only on divergence in proportional terms: that is, assessing the extent to which the digitised corpus is a micro version of the larger population (by proportionally adhering to the contours of the landscape). However, it needs to be noted that the press itself was skewed towards the more privileged groups in society, and a proportionally representative sample, if achieved, would still incorporate these biases. For some research

scenarios it might be preferable to compute bias by comparing the sample to a distribution that gives *equal* weight to different key characteristics such as political allegiance, price or place of publication even though this was not the distribution in the historical landscape of newspapers.

This is perhaps one of the most exciting opportunities of the Environmental Scan: it gives us the opportunity to capture and foreground different worldviews. Klein et al. argue that

> [r]eframing bias not as 'bad data' but as an effect of unequal structural power redirects our attention to the social, historical, and political conditions that gave rise to the biased data, as well as to the spaces of indeterminacy, and to the irrecoverably missing parts, of any dataset past or present.[55]

The Environmental Scan is a method for documenting and naming some of the attributes that contribute to, or act as proxies for, these social, historical and political conditions. But documenting is also an enabling act. By creating enhanced metadata labels, we give users the ability to create their own digital subsample of datasets that meet their own set of criteria. In this way, the Environmental Scan goes beyond a passive measuring process to facilitate a more active and experimental approach to studying the past. Researchers can construct corpora that reflect alternate worldviews by weighting their data for different attributes: a subsample that perfectly balances liberal and conservative voices; one that focuses only on daily publications at the cheapest price points; or newspapers claiming to represent the working classes, and so on. Such an approach could thus be used to recentre voices from historically marginalised, underrepresented perspectives (as long as *some* newspapers sought to reflect this perspective at the time).

In LwM we developed a tool that facilitates the alternating of worldview by creating collections of newspaper data based on specific sampling criteria. LwM research software engineer Tim Hobson created a Python tool *subsamplr* (which can be found among the additional resources in this book) to help researchers build informed subsamples of the newspaper data.[56] Subsamplr offers various sampling strategies for constructing corpora that adhere to a predefined shape. More specifically, it enables prescriptive samples, obtained by random selection while prescribing a desired distribution with respect to particular variables.[57] The combination of enriched metadata and sampling offers scholars a novel way to

understand better how the shape of the corpus influences research outcomes or to check for the robustness of their findings. Another application relates to machine learning and more specifically language modelling: does the preponderance of particular worldviews in the data propagate as biases in the model itself? Or put another way, to what extent does ensuring the equal representation of voices create 'fairer' artificial intelligence?

Extrapolating the deep and wide scans: working with maps and books

We have outlined our framing principles and our operational steps in the Environmental Scan through the example of the digitised newspapers and press directories, alluding in the process to where these principles and steps might be generalised. In the following sections we outline how these principles can be adapted to two other important collection types – maps and books – two areas where we believe there are particular opportunities for at-scale source criticism. Those more interested in the general principles of the Environmental Scan may prefer to jump to the chapter conclusion.

Scanning maps

Periodicals like newspapers are often catalogued at the title level rather than a more granular record (e.g. individual issues), and series maps are similar. Series maps are multi-sheet sets that, laid out together, map a continuous part of the earth, often for a nation state. United States Geological Survey maps, French Cassini and *Etat Major* maps, and British Ordnance Survey maps are well-known examples and, like newspapers, are media whose publication ballooned in the nineteenth century because of new printing techniques and state, commercial, educational and public interest in maps.[58]

Because the number of sheets in even one full set of one such national survey can jump into the thousands, series maps usually fall into the class of collections that libraries rarely have resources to catalogue on an item-by-item basis. And yet, at their time of publication, series maps were novel, complex publications that required new tools for navigating among the sheets: index maps were created to accommodate this need (Figure 1.6) and are often used by libraries to aid in finding sheets for readers in the absence of catalogue records. If individual sheets represent a level of detail of the

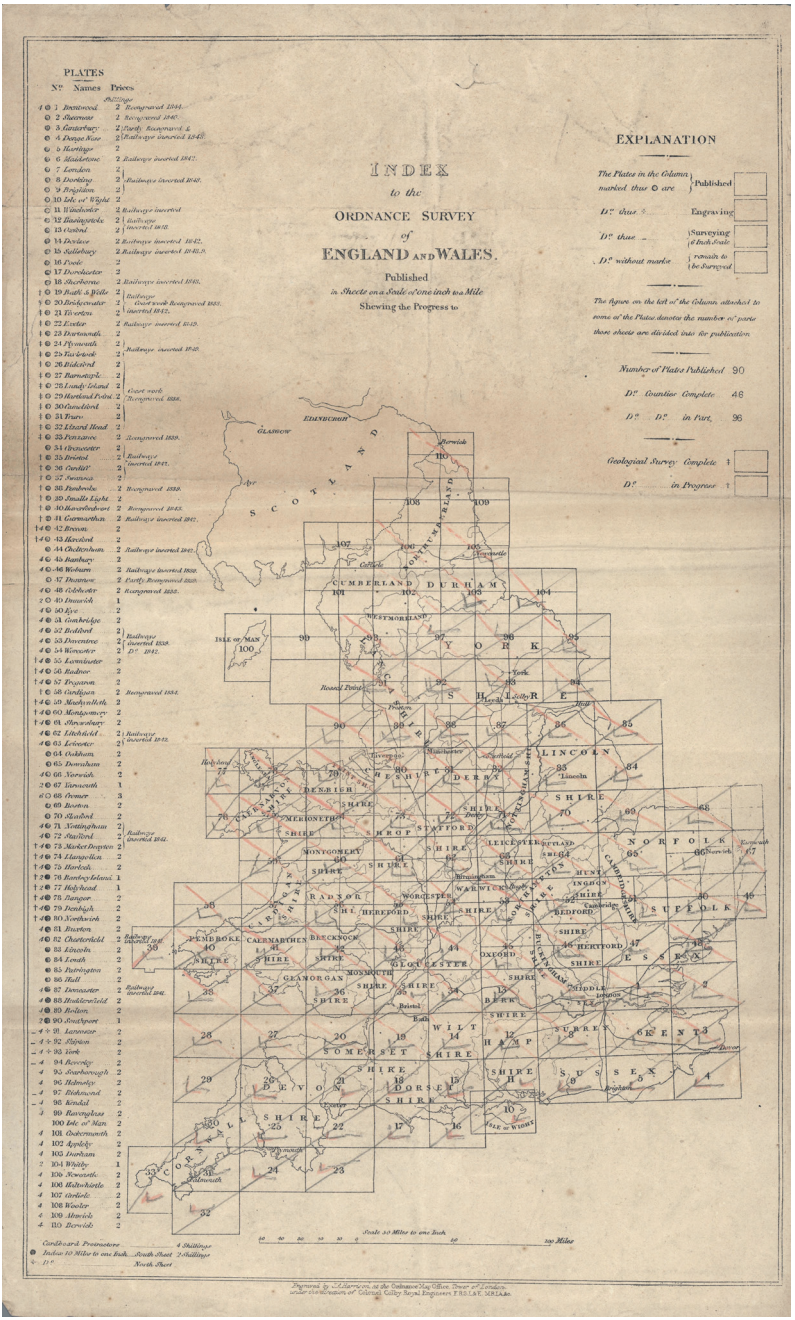


Figure 1.6: Ordnance Survey of Great Britain, William Mudge, Thomas Colby, Benjamin Baker, *Index to the Ordnance Survey of England and Wales, Published in Sheets on a Scale of one inch to a Mile* (Ordnance Survey, 1870).[59] Image provided courtesy of David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries. CC BY-NC-SA 3.0.[60]

then industrialising landscape that had never been attempted before, index maps are the means by which this work was synthesised. In Environmental Scan terms, the index map provides a framework similar to the *Directory* structure. It was created to communicate about the abstract national landscape being sold sheet by sheet, a powerful message that was also a handy finding aid. For scholars today, the concept of the index map serves as the foundation against which to explore bias in physical and digital collections, and it's no surprise that the most intuitive user interface for searching digitised map collections like the OS is a web-based index map (Figure 1.7).

Over time, individual sheets in series map sets have been much used by historians and other scholars *as reference sources*: they are handy for finding places or features that scholars are already aware of, confirming where things are with more precision, or adding spatial context to one's understanding of the landscape or city. But they have rarely been used *as collections* whose expansive geographical coverage of regions and nations permits systematic searching for map elements that are not predetermined. Efforts to explore the history of the national mapping projects that created these nineteenth- and twentieth-century series maps have resulted both in narrative histories about their production

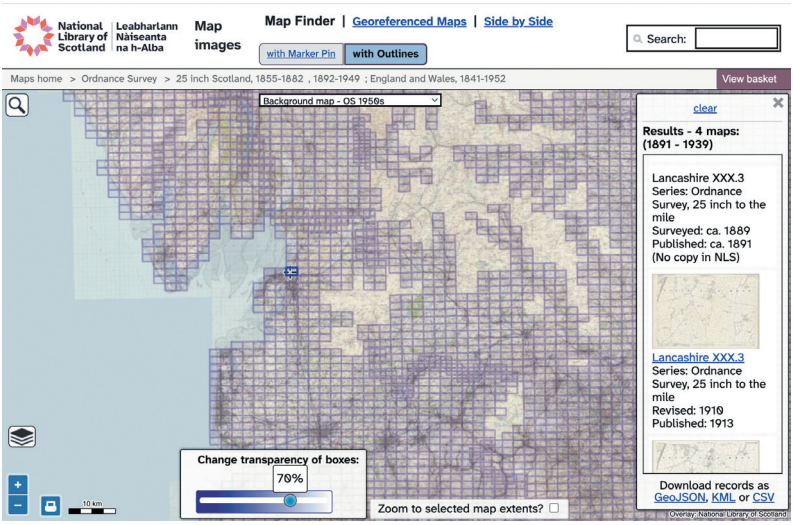


Figure 1.7: National Library of Scotland's 'Map Finder' interface.[61]

as well as detailed carto-bibliographies for the more manageable collections.[62] If carto-bibliographies take a holistic approach to map production and tracking extant copies in collections, there has been little investigation of the content of these maps at the collection level. The 180-sheet national 'Cassini' map of *ancien régime* France (see Figure 1.8) is one of the rare examples of a series map that has been almost completely 'digitised': large projects in France like *GéoPeuple* and its successor *GeoHistoricalData* have created datasets of symbols and other features printed on Cassini sheets enabling their use as data for historical research.[63] OS map sets of



Figure 1.8: Louis Capitaine, César-François Cassini, and the Cassini family, *Tableau de la carte générale de la France* (L. Capitaine, 1797) in César-François Cassini and the Cassini family, *Carte de France* (L. Capitaine, 1750–1815).[64] Image provided courtesy of David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries. CC BY-NC-SA 3.0.

nineteenth-century Britain are, however, much larger than the Cassini map, with the six-inch scale coming in at almost 16,000 sheets per edition, and the larger-scale twenty-five-inch scale numbering around 50,000 sheets per edition. Any similar manual digitisation of OS features from the historical maps has been unfathomable given the extraordinary number of sheets.

On LwM, we sought to overcome this challenge by creating data from OS map content using machine learning rather than traditional, manual GIS methods.[65] We began with the OS maps recently scanned and georeferenced by the NLS. Later in the book we talk about MapReader, the computer vision software library we developed to analyse this map data at scale (Chapter 2), as well as how the outputs of this software – combined with other datasets including census returns – allowed us to understand the impact of the arrival of rail on different communities (Chapter 3). But before we could do that, we needed to learn how to think historically with whole collections of maps. To apply the principles of the Environmental Scan to OS maps is to consider how to look at sheets in the context of their creation (the deep scan) and in relation to the 'landscape' of associated maps, building on the concept of the index map (the wide scan). With maps, the Environmental Scan adopts a processual approach to map history that illuminates the practices and ideals shaping each map as well as an entire series as an essential element in interpreting their content.[66] For example, survey, revision and print dates shatter any illusion of temporal unity for the different nationwide OS map editions: the second edition of the six-inch-to-one-mile maps took almost forty years to complete. As with newspapers, we propose that understanding the shape of a collection is a prerequisite to both formulating and answering meaningful research questions.

Map curators internationally are fully aware of the challenges around uncatalogued map collections. If processing individual items failed at the moment of acquisition, this opportunity presents itself again during digitisation. Despite the financial constraints around digitisation, some series map collections are emerging with at least minimal item-level metadata linked to scanned images of maps.[67] We were lucky that the NLS funded the creation of metadata for the entire OS 'archive' of maps, regardless of whether the physical sheets were held by them.[68] Akin to the *Newspaper Press Directory* content, such metadata provided key information for a 'wide' scan, allowing us to understand the relationship between what had been digitised, what was held by the NLS, and what had been originally

printed by the OS historically. The NLS Maps website encourages exploration of their collections and makes use of this metadata and additional enrichments to allow users to search by a range of metadata fields, including georeferencing information linked to gazetteers that enable spatial searching. The 'Map Records search' service is the closest thing to an experience of catalogue searching for OS maps (Figure 1.9).[69] It's possible, for example, to search spatially by modern place names (e.g. navigate to maps that contain the latitude/longitude coordinates of a given place), or by latitude/longitude, National Grid Reference or Post Code directly. From a collection-as-data perspective, visualising the NLS 'footprint' of specific series (all sheets in a series, across editions) as it is preserved in the twenty-first century creates the impression of a 'national' map. This visual interface was an important step in presenting digitised map collections to the public: the NLS was among the first libraries in the world to make collections discoverable in this way, setting a foundation for the 'wide' Environmental Scan. For many users, this 'browsable' search interface is sufficient. But, as this exploration of the digital Environmental Scan demonstrates, it is essential to be able to move from a browsing mode to a 'critical search' mode.

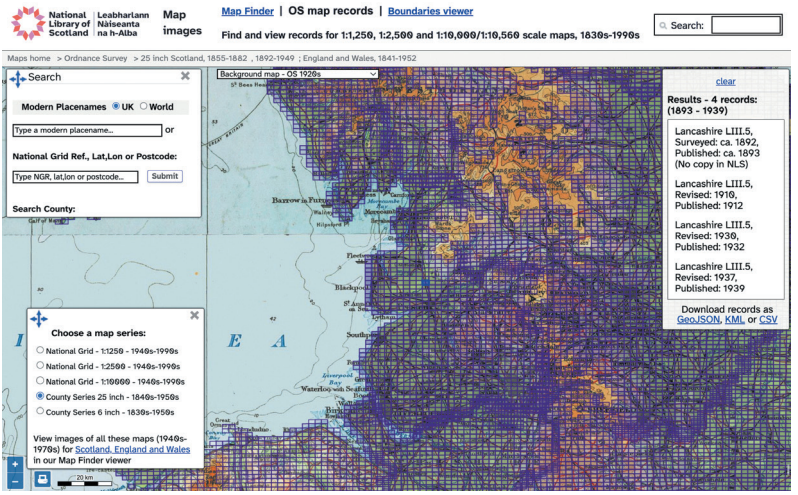


Figure 1.9: The National Library of Scotland 'OS Map Records Search' interface showing 25-inch sheet bounding boxes, with specific sheets printed over time and available in the NLS collections displayed in the top right-hand corner.[70]

Enable critical search

'Critical search' is a term coined by Jo Guldi to describe an iterative workflow or 'procedure' of information gathering where the user actively analyses and evaluates the credibility, bias and context of the sources they find.[71] In order to work in a digital environment with these maps in such a way, we needed both the map images themselves and their associated item-level metadata (shared as GIS shapefiles). Beyond viewing maps on the website, NLS makes several of their editions and scales of OS maps available through tileservers.[72] This approach allows the sheets of adjacent maps to appear as a single, stitched-together 'composite map', which in the case of some of these NLS collections, covers most of Scotland, England, Wales and Ireland.[73] NLS uses cloud platforms to provide composite maps as tile layers, which can be downloaded at specific zoom levels. In conjunction with the NLS OS metadata, it is possible to specify which 'sheets' (e.g. tiles within a geospatial bounding box that replicates the scope of an original printed sheet) to download.[74] As with newspapers, we needed not only to understand what was in the NLS's physical OS collection and its history but also to clarify what was presented in each digital version of each scale over time. For this, we turned to exploratory data visualisation to tackle the 'wide' part of the Environmental Scan.

In keeping with the project's central ethos, we wanted not only to solve a problem for ourselves (and about nineteenth-century history) but also to create resources that could have utility beyond the project and model this exploratory first step essential to digital research with serial collections. Despite a rich tradition of writing about the OS, there was no simple way to compare what had been printed with what had been digitised. With Olivia Vane, a data visualisation specialist on the project, we developed new approaches to displaying metadata, map sheet bounding boxes and map image thumbnails as basic components in exploring the NLS OS map collections. First, we needed a simple solution to see different scales of maps as they were published over time (in multiple editions and revisions): OS maps at 1-inch, 6-inch and 25-inches to 1 mile were most relevant for us. We needed a visualisation that would help us triangulate between three connected problems: (1) what has been digitised; (2) what kind of content does each scale depict; and (3) what is significant historically?

In practice, this first visual step in the 'wide' Environmental Scan of OS maps allowed us to have a bird's-eye view of what maps are present

for different periods. Taking inspiration from historical OS index maps and the NLS website visual indexes,[75] we developed the small-multiples-inspired 'Macromap' visualisation.[76] Users can zoom in and out of the different scales (displayed horizontally), customise the time steps (displayed vertically), and add details like individual sheet outlines or colour for non-digitised sheets (Figure 1.10). Just as newspaper metadata helps pinpoint particularly rich periods where many titles have been digitised, Macromap provides a dynamic workspace for comparing very big map collections in one view. This simplifies answering questions such as: what regions should I focus on if I need to look at how the landscape is changing over time (e.g. across editions)? Which scales and editions have the richest content from the 1880s? Which scales and editions would I prioritise if I want as much of a three-nations (England, Wales and Scotland) dataset as possible (e.g. those that have the fewest gaps in the NLS' digitised

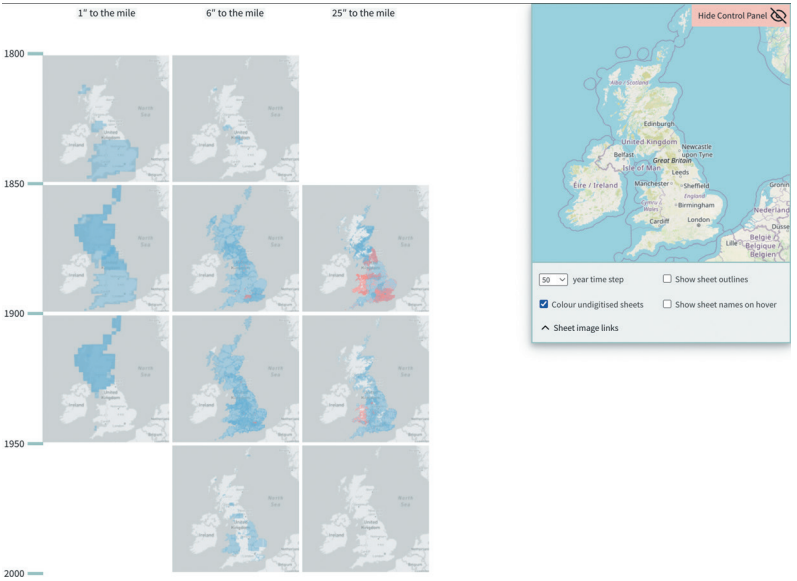


Figure 1.10: Detail from Macromap of three scales of OS maps plotted over time based on their earliest available survey or print date. Blue shading represents sheets that have been scanned by the NLS. Red shading represents sheets that were printed but have not been scanned by the NLS. The interactive version of this figure allows you to navigate within the Macromap, while the static version shows a screenshot of the full-screen Macromap interface, including the control panel. Credit: Olivia Vane.

collections)? Finally, the zoom feature means that it's possible to view maps at this multi-national scale and to 'winnow down' (a step in Guldi's procedure) to a single sheet and see a hyperlink that takes you directly to the NLS map viewer for that sheet (Figure 1.11).

Engage with the complexity of temporal change

Macromap highlights the complex temporalities of the OS. As discussed above with the data linking process of the *Newspaper Press Directory*, as well as longitudinal analysis of the bias of the digital newspaper collections, temporality introduces considerable complexity into any Environmental Scan. But the strategies of the Scan – the process of adding enhanced metadata and the affordance of dynamic visualisation – enable us to grasp these complexities in ways that have not been possible heretofore. In order to have a clearer view of the chronology of sheets within one edition of one scale of the OS maps, Vane developed 'animated map chronologies' as Observable notebooks (Figures 1.12 and 1.13).[77] These visualisations have been powerful reminders that OS editions were the product of multiple decades of work.[78] Seeing the thumbnails for each sheet appear based on their year of survey is almost like following the surveyors themselves around the landscape. These animations are a stark reminder that the

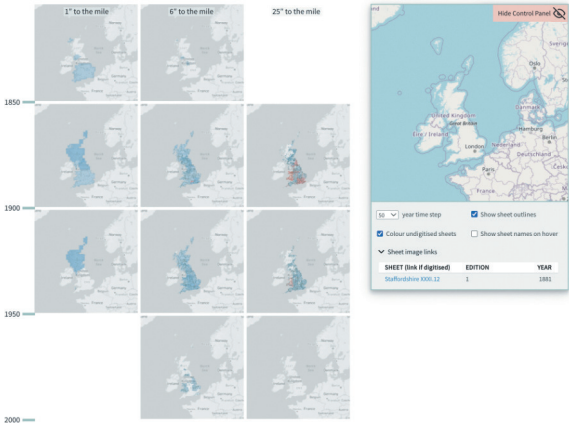


Figure 1.11: Detail from Macromap showing sheet link in the bottom of the Control Panel, which appears after selecting the sheet's location on one of the mini maps to the left. Credit: Olivia Vane.

maps are not one giant national map, but a puzzle of sheets that appeared, in the case of the second edition six-inch maps that we worked with the most, over almost forty years. Furthermore, these maps reappeared over time as they were revised and reprinted, something that the OS metadata only rarely captures.[79] As tools for communicating about the complexity

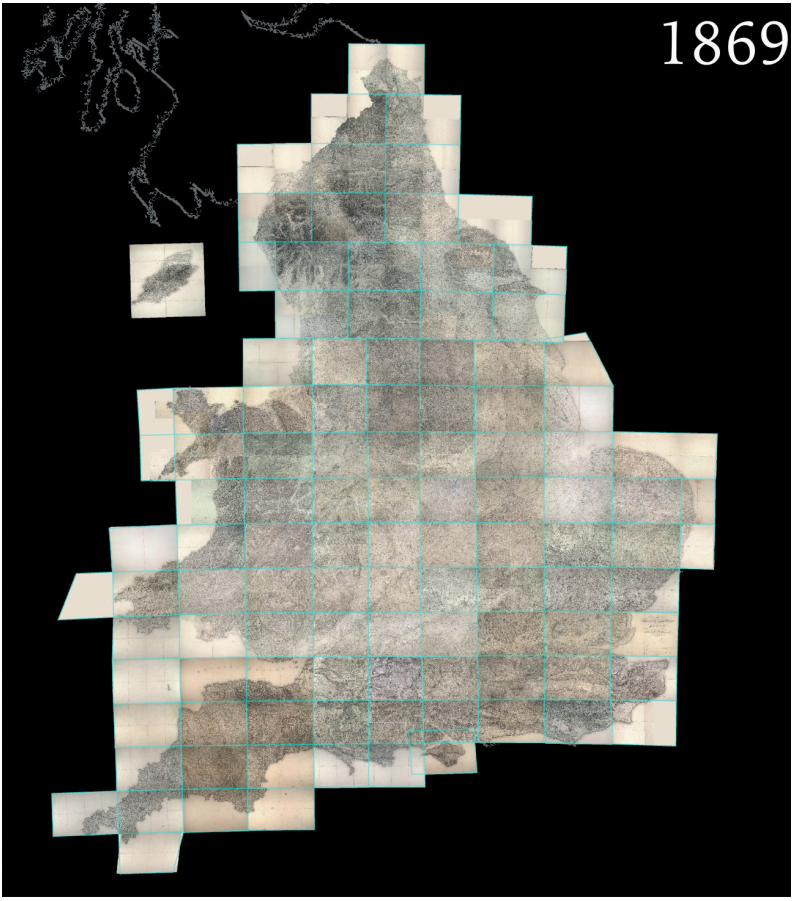


Figure 1.12: Olivia Vane, *OS 'Old Series'/'First Series', England & Wales* (2023).[80] The interactive version of this figure shows the Old (or First) series sheets appearing across the map according to the year they were printed. The static version for print shows one screenshot depicting the sheets printed by 1869. Credit: Olivia Vane. Map metadata courtesy of the Charles Close Society. Map images courtesy of the British Library (see note 73).

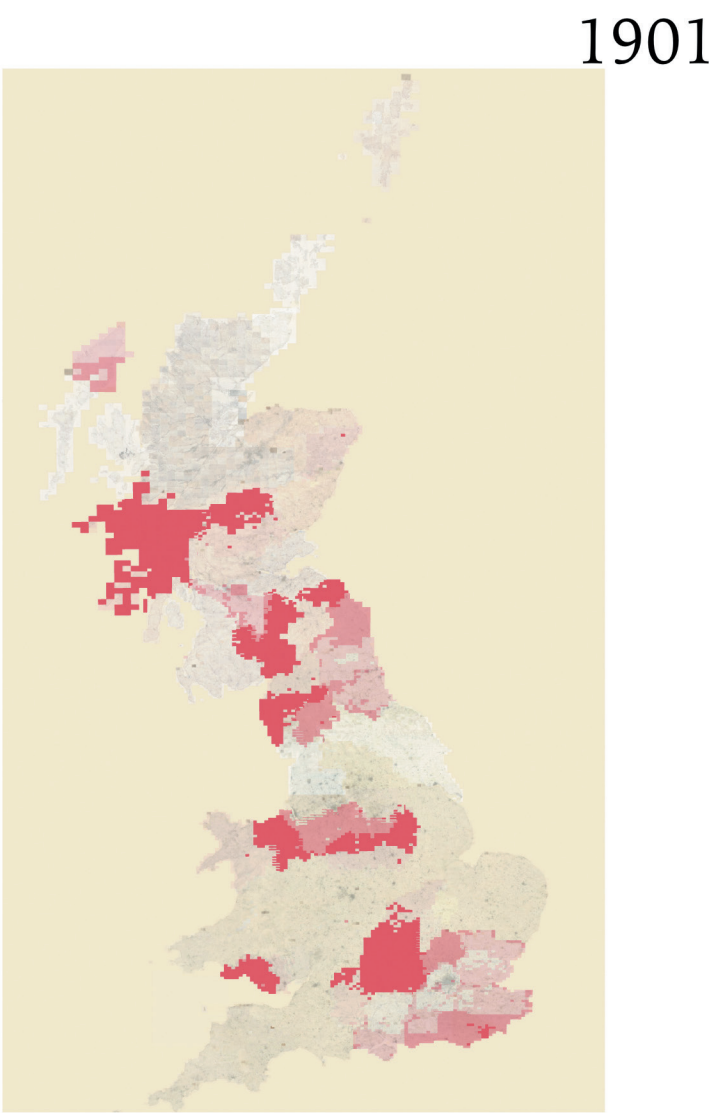


Figure 1.13: Olivia Vane, *OS 6"-to-the-mile, Great Britain: 2nd Ed.* (2021).[81] The interactive version of this figure shows the six-inch sheets appearing across the map according to the year they were printed. The static version for print shows one screenshot depicting six-inch sheets printed in 1901. Credit: Olivia Vane. Map images and metadata courtesy of National Library of Scotland.

of visual collections, like maps, data visualisations have been far more effective than other forms of exploring the metadata, such as tables listing the numbers of sheets printed and those that were digitised, by scale and edition. Macromap and the animated map chronologies render map metadata into spatial interfaces that historicise digitised collections that can too often be collapsed into data divorced from their physical origins.

The Environmental Scan of OS maps enabled by these tools was an essential step in our research process, and one that allowed us to work closely with librarians who curate the OS collections at the NLS and the British Library. This is an important point about the human factor embedded in the digital Environmental Scan approach: while it is fundamentally a set of data processing, linking and analysis tasks, it is based on conversations and collaborations between curators who maintain reference collections of resources like press directories or collection metadata. We cannot emphasise enough how much our understanding and use of digital collections is dependent on the expertise of colleagues working in libraries and archives.

Find alternative ways to contextualise data

For newspapers, the 'deep' and 'wide' scans both contribute to the informed selection of a set of titles, issues or items for analysis. For maps, we might think of these steps as sequential. After using the 'wide' scan to aid in selecting a set of maps for an experiment, the next step is to understand the relationship between what is on the map, why it is on the map, and what it might be possible to say about this map content at scale. This version of the 'deep' scan for maps is slightly different because, at the outset of LwM, we did not yet have the equivalent of full-text content to analyse from maps. Later in this book, we review the work to develop MapReader, a software library which creates structured data from map content (see Chapter 2). What would be the equivalent of the press directories for maps? Rather than one reference source playing this role, for maps there will be many answers, depending on the map content a researcher is interested in. For example, in our work on identifying railway infrastructure on the six-inch maps, we used a dataset of historical railway passenger stations, which included metadata about which companies operated out of each station, to complement our understanding of which companies were likely to be involved in the creation and maintenance of adjacent railways (see Chapter 3).[82]

New opportunities for using both historical and present-day data as a reference to enrich specific map content identified with MapReader will mean that we are able to refine our ability to date the revision of OS map sheets ('states') beyond the information now available in the NLS metadata. For example, we have developed methods for detecting and transcribing text on maps, which can then be linked to external knowledge bases like Wikidata or World Historical Gazetteer.[83] This new metadata complements the content, as with political leanings and prices for newspapers. In the future, we may add to the existing metadata with information about when and where specific sheets were sold, what libraries continue to have holdings of specific states (the carto-bibliographic style approach), what prices were charged for sheets, how many sheets of each state were printed, and more. From the OS archive to library catalogue data worldwide, there are many possibilities for the deep scan of this particular map series at the heart of modern British history.

The approach to a digital Environmental Scan of OS maps – from the library infrastructure that enabled it to the tools that we built on LwM to complement that foundation – is one that we like to think can be replicated with other big map collections. The building blocks can take time to assemble and will likely require collaboration with map curators and metadata librarians, but we offer the digital Environmental Scan as a starting point for digital source criticism of these usually vast collections. If computational research with other series map collections around the world takes place against the backdrop of the Environmental Scan approach, it will be a big step forward in bringing historical methods to bear on the use of big cartographic collections.

As more historians work with large datasets, and as we develop very different research practices, we require this investment in data exploration and metadata enrichment as a way to know our sources, the contexts of their production and reception, and their curatorial histories. Foregrounding this history of data as a core part of any computational research project is at the heart of the digital Environmental Scan. This is perhaps nowhere more pertinent than the study of large-scale book corpora.

Scanning books

Book corpora have often been equated with representing 'culture' more broadly, especially in the Culturomics traditions, which leverage very large

but still specific sources of information as they are based principally on university library collections.[84] In this final section we focus on a popular collection of open access nineteenth-century books from the British Library collection. Because of its vast scope this collection might entice researchers to use it as a proxy for 'Victorian literature' or even 'nineteenth-century culture', however, as we argue below, we should tread carefully and at least attempt to qualify such claims in more detail.

Illuminate obscure selection processes

In 2005, the British Library partnered with Microsoft to digitise a selection of out-of-copyright texts from the eighteenth and nineteenth centuries. This collection, referred to throughout this document as 'BL Books' or 'Microsoft Digitised Books', was made openly accessible in various formats, including images, XML and JSON with plain text – the latter being the version we used for this research. The collection comprises approximately 49,455 titles (about 65,227 volumes) published between 1789 and 1914. While this sounds like a large quantity of books, we established that the Microsoft Books collection constitutes just 2.76 per cent of the whole British Library catalogue (only looking at publications dating from the nineteenth century).

Moreover, the usefulness of the corpus is somewhat affected by the lack of transparency around its composition, and what this might mean for representativeness. The history of the digitised collection remains obscure, as the selection criteria for digitisation are unknown. What we do know is that this digitisation project emerged in competition with Google, which, around the same time, began rolling out its Books program in collaboration with various university libraries worldwide. Consequently, we are confronted with an 'interrupted' dataset, frozen at a particular point in the digitisation process.

To understand the effects of this (random) freeze on the shape of the digitised collection, we developed the 'Sample Inspector' as an implementation of the wide scan. The Sample Inspector is a series of Jupyter notebooks now released as part of the project 'Classifying 19th Century British Library Books using Crowdsourcing and Machine Learning'.[85] Through these notebooks, we investigated whether the collection of digitised titles mimics the known 'landscape' of book publications in the nineteenth century. To do this we compared the relative frequency of keywords in titles between the corpus of digitised titles and the complete holdings of the British Library

(as a surrogate for the broader publishing landscape). The British Library data services provided the latter information in JSON format.

Given that digitisation likely proceeded shelf-by-shelf and was halted before completion, we might expect substantial gaps regarding topics or genres. Sample Inspector is primarily aimed at understanding topical lacunae in the digitised corpus. This is especially crucial when a scholar, computational or not, is interested in using this large corpus to explore a specific topic or type of literature. For example, during the LwM project we used the book corpus to study the trope of 'living machines' in the nineteenth century. We needed to understand to what extent this collection might under- or over-represent literature related to machinery (based on the presence of 'machine words' in book titles). As Figure 1.14a shows, there were clear differences between the digitised corpus (blue line) and the wider landscape of book holdings in the BL for our 'machine' keywords (orange line). Compared to the whole collection, few books on machines were digitised and the evolution over time

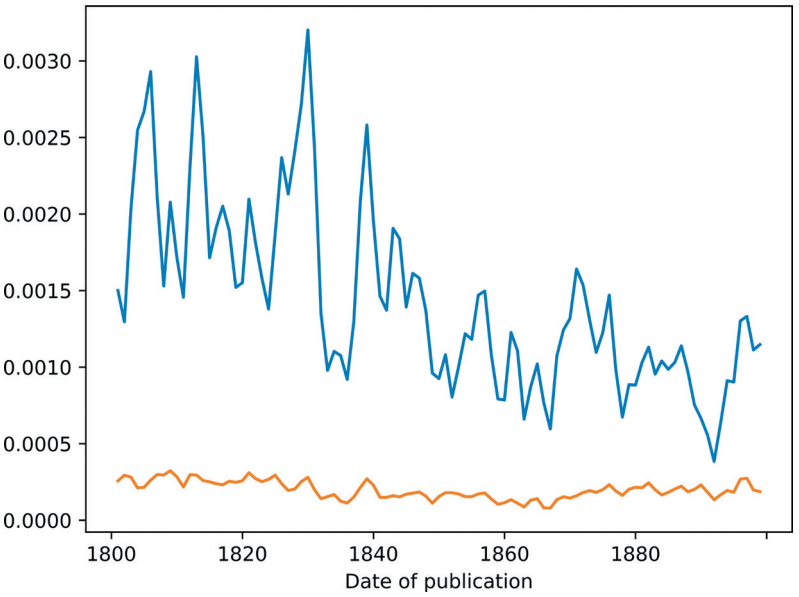


Figure 1.14a: Relative frequency of the words 'travel(s)/travelling' in book titles. Sample represented by a blue line and population by an orange line. Credit: Kaspar Beelen.

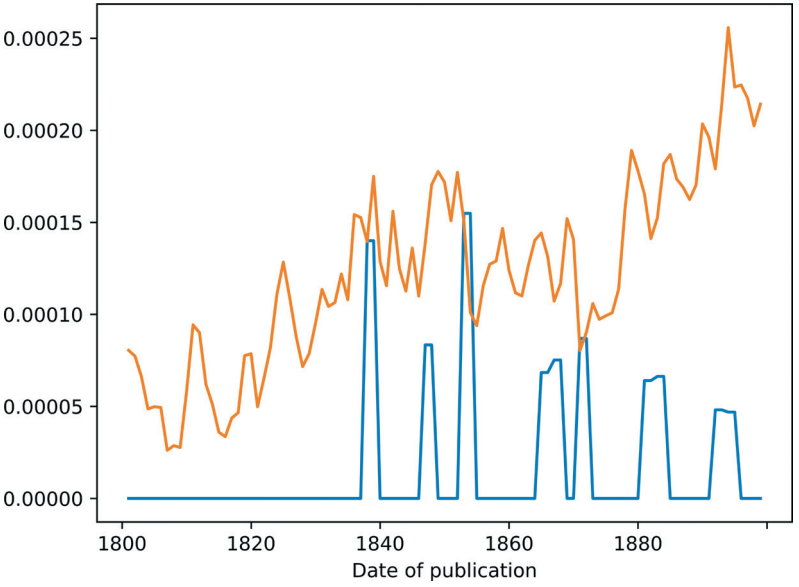


Figure 1.14b: Relative frequency for the words 'machine(s)' and 'engine(s)' in book titles. Sample represented by a blue line and population by an orange line. Credit: Kaspar Beelen.

doesn't follow the overall historical pattern based on the BL metadata. When examining the results for other queries, the biases induced by the opaque selection became even more apparent. The digital sample appears to radically oversample titles on 'travel' (blue line, Figure 1.14b) compared with the wider, undigitised library collection (orange line). Knowing more about the wider landscape of materials for which we lack digital access is critical if we want to make sense of findings built on such digital resources.

Generate new metadata through annotation and classification tasks

Along with gaps in topical coverage, we also explored the issue of genre to assess the extent to which the digitised corpus was skewed towards fiction or non-fiction. However, the metadata associated with these books is inconsistent and incomplete. British Library Labs aims to release subsets of the collection – such as all poetry or publications from specific countries – as research datasets. Unfortunately, the current lack of high-quality metadata

currently makes this infeasible at present. To enhance and expand existing metadata, Victoria Morris from the British Library's Metadata Standards team launched a crowdsourcing project on the Zooniverse platform. The project's goal was to enrich the records of Microsoft Digitised Books by gathering information on the following aspects:

- country and place of publication
- date of publication
- edition
- language(s) of the content
- literary form (fiction or non-fiction)
- genre (for fiction items)
- subject.

For the purposes of this project, the term 'literary form' had a specific scope: 'fiction' was broadly interpreted to include all creative writing, such as poetry and drama as well as novels and stories. This definition may differ from how 'fiction' is understood in other contexts, particularly when identifying narrower or more specialised genres.

This crowdsourcing project made a valuable contribution toward expanding the metadata coverage of the collection. However, it was only able to update the metadata for a fraction of the entire collection as shown in Figure 1.15 – which shows the number of records which have been annotated (i.e. checked, listed here as 'true') against those that remain to be annotated ('false') – and led us to consider the potential role of machine learning for closing the annotation gap. We aimed to explore whether a model could be trained to apply the metadata generated

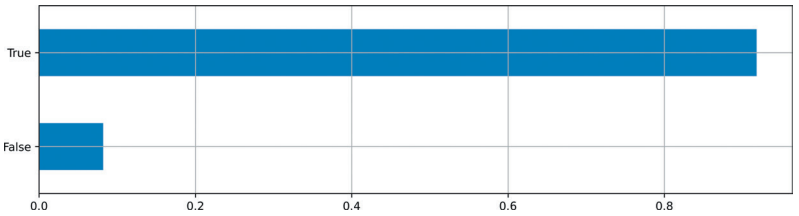


Figure 1.15: The proportion of metadata annotated ('true') versus not annotated ('false') in the BL Microsoft Books collection. Credit: Daniel van Strien.

through crowdsourcing to the entire collection. Specifically, we wanted to investigate if the model could determine whether a book was fiction or non-fiction based solely on its title. With such a model, we could not only expand the metadata of BL Books but also classify all library catalogue records, thereby analysing the divergence between the digitised corpus and the British Library's holdings with respect to genre.

The genre classifier consisted of a distilbert-base-cased language model (a smaller variant of BERT with lower computational requirements) fine-tuned on annotations from the Zooniverse crowdsourcing tasks, which consisted of more than 4,400 labelled examples.[86] The labelled dataset was expanded by applying 'labelling functions' to other titles from the Microsoft Books Corpus. These labelling functions automatically expand the set of annotations based on rules (or more complex algorithms). A simple labelling function maps a title to labels based on certain words. For example, we can easily expand the number of annotations by automatically classifying a book as 'fiction' if the title contains the word 'novel'. We created similar labelling functions based on keywords (e.g. words such as 'biology' strongly correlate with 'non-fiction'), named entities (e.g. titles with organisations in them are labelled as 'non-fiction') and other deductions drawn from user experience with the dataset (e.g. we assumed long titles are mostly 'non-fiction' works).[87] Based on these simple rules we generated 34,542 additional labelled examples on which to train the genre classifying model. The eventual evaluation accuracy score on the test set is 0.95 – meaning that out of a hundred predictions ninety-five are correct – a satisfactory result for such a complex task.

Table 1.3: Evaluation results for the fiction/non-fiction classifier

	Precision	Recall	F1-score	Support
Non-fiction	0.91	0.94	0.93	296
Fiction	0.97	0.95	0.96	554
Accuracy			0.95	850
Macro avg	0.94	0.95	0.94	850
Weighted avg	0.95	0.95	0.95	850

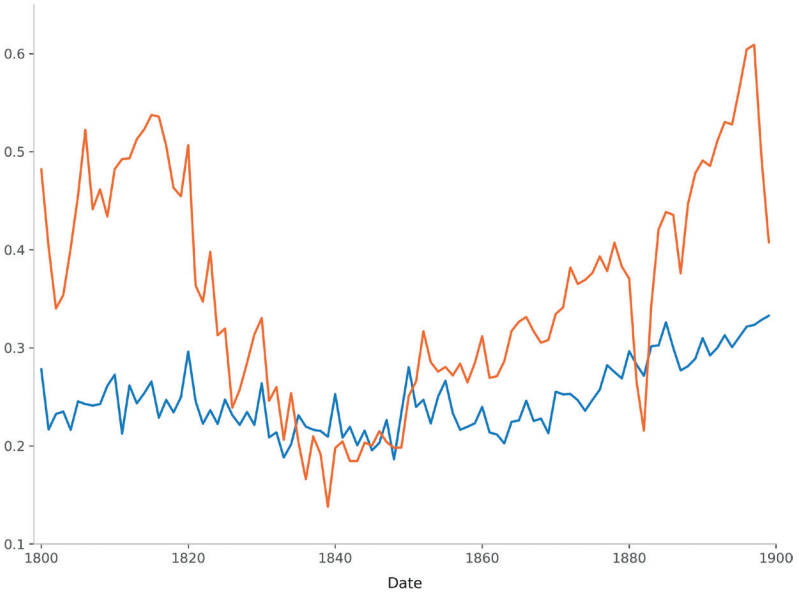


Figure 1.16: The proportion of fiction works in the Microsoft Books Corpus (orange) and British Library Catalogue (blue). Credit: Kaspar Beelen.

After training and evaluating the classifier, we applied it to both metadata records, labelling each title, and then computed the portion of fiction works by year. Figure 1.16 clearly shows how, with respect to genre, the sample and population of books diverge over time, with the digital sample over-representing fiction titles significantly in the early nineteenth century and for most years after 1850. Only in the 1830s and briefly in the early 1880s is fiction under-represented. This may be by chance, or it may be because these were famously two periods when social questions, the so-called 'condition of England' debate, dominated public discourse, potentially influencing curators' early decisions about what most needed to be digitised.[88]

Conclusion

There can be no simple, universal answer to the problem of how to do source criticism at scale. This is due to the diversity of digital sources and

the myriad research questions that can be addressed when using them. However, the case studies discussed in this chapter illustrate a set of key principles and strategies that can guide the design of digital Environmental Scans that fit the challenges of different collection types. Solutions and implementations depend on contexts, research questions and access to sources.

The kinds of diagnostic tools that we have described enable users to home in on issues such as representativeness and bias in historical collections. Our aim in this context is not to suggest we can 'solve' or 'discard' bias but rather to enhance transparency with respect to the composition of digital research materials. Most importantly, the Environmental Scan method helps users assess which voices remain under-represented in the data. Not only does this inform our research but it can also contribute to building a more equitable historical collection by guiding future digitisation. For example, based on the enriched metadata, we can identify and prioritise digitising newspapers that advocate for the interests of groups marginalised by class, ethnicity or gender. The presence of such sources in the digital archive could then be boosted to better reflect these groups' presence in society at large, if not in the public sphere.

If continued digitisation is not an option, the scan makes it possible for users of a collection to design research questions and methodologies that take these previously invisible archival contours into consideration. Alternatively, it can also enable researchers to construct custom subsamples of the data that meet their own criteria for research, including creative and experimental approaches that would allow them to construct samples reflecting particular worldviews. Based on information obtained via a 'deep scan', researchers could artificially construct corpora adhering to specified constraints, for example, by oversampling sources from marginalised voices or perspectives and downsampling those arising from more privileged and dominant groups. Such an approach does not aim to 'correct' for bias but can provide a critical tool to investigate how the historical and social inequalities that have shaped collections (and continue to do so today) inform the results we obtain when using digital methods uncritically. We are actively pursuing some of these developments in our own research and hope that the philosophy behind the Environmental Scan will encourage others to do the same.

While newspapers, maps and books provide an excellent proof-of-concept to demonstrate the core principles of this approach, these

methods have wider applicability. Crucially, not only can the Environmental Scan be developed to process diverse inputs (i.e. different types of digital collection) but it can also yield outcomes with wide-ranging benefits. Most importantly, the Environmental Scan can be broadly understood as a diagnostic toolkit that assists in identifying various types of unintended distortions that can creep into large heritage data. In this respect, the scan has the potential to support and strengthen myriad research findings and applications that build on these immense collections of historical data.

Addressing these issues will not only advance digital scholarship of the past, but it might also persuade sceptics that 'distant reading' by machines can be supported by source criticism at scale and therefore does not have to take place in a vacuum where data is stripped of historical context.

But, if we are bold, perhaps we can suggest the payoff is bigger still. Here there is a particular opportunity for humanities researchers and GLAM organisations to show the important perspective they have to offer in the current moment of AI-hyperbole. Historical collection data represents the perfect testbed for showing that 'bias' is necessarily baked into any collection of cultural products: they can be analysed to show the different situated perspectives that they encode,[89] and how these are anchored to specific moments in time, specific places and communities of knowledge, as well as 'to the spaces of indeterminacy, and to the irrecoverably missing parts'.[90] The Environmental Scan is a set of principles for making that complex history legible to machines and humans alike and its principles can and should be extended to the scraped data that underpins the large language models currently driving the AI revolution.

Notes

1 Beelen and Lawrence conceptualised the paper. Beelen and Lawrence drafted the first two sections (focusing on Newspapers). Lawrence wrote the first draft of the conclusion. The materials of this chapter are based on collaborations between Beelen, Lawrence, Wilson and McDonough. McDonough wrote the 'Extrapolating the deep and wide scans' section (focusing on Maps) based on collaboration with Olivia Vane and the LwM team. Beelen wrote the 'Scanning books' section based on the work and blog posts by van Strien. Ahnert reorganised the draft and developed the present argumentative structure of the paper. All authors contributed to draft revisions and to the conclusion.

2 See the user discussion at: https://en.deezercommunity.com/ideas/who-is-the-composer-improved-labelling-of-music-551 (accessed 17 September 2024).

3 Lara Putnam, 'The Transnational and the Text-searchable: Digitized Sources and the Shadows they Cast', *The American Historical Review*, 121, no. 2 (2016): 377–402.

4 This took place around 2012 and the database mentioned was: https://plenum.be/ (accessed 15 June 2026).

5 More broadly on the perils of keyword search see, Ian Milligan, *History in the Age of Abundance: How the Web Is Transforming Historical Research* (McGill-Queen's University Press, 2019), 45, 56–8.

6 Tim Hitchcock, 'Confronting the Digital: Or How Academic History Writing Lost the Plot', *Cultural and Social History* 10 (2013): 9–23, https://doi.org/10.2752/147800413X13515292098070; Ian Milligan, *The Transformation of Historical Research in the Digital Age* (Cambridge University Press, 2022).

7 Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (University of Chicago Press, 2019).

8 Daniel C.S. Wilson 'Working at Scale: What Do Computational Methods Mean for Research Using Cases, Models and Collections?' *Science Museum Group Journal* 18 (2023), https://doi.org/10.15180/221805.

9 Laurel Brake, 'Nineteenth-century Newspaper Press Directories: The National Gallery of the British Press' *Victorian Period Review*, 48 (2015): 569–90, https://doi.org/10.1353/vpr.2015.0055. The earliest and longest-running publication was Mitchell's *Newspaper Press Directory*, begun in 1846 and published annually from 1856.

10 Kaspar Beelen, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan', *Digital Scholarship in the Humanities* 38, no. 1 (2023): 1–22, https://doi.org/10.1093/llc/fqac037.

11 Rob Kitchin, 'Big Data and Human Geography: Opportunities, Challenges and Risks', *Dialogues in Human Geography* 3, no. 3 (2013): 262–7 and Rob Kitchin, 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society* 1, no. 1 (2014): 1–12, https://doi.org/10.1177/2053951714528481.

12 Kitchin, 'Big Data, New Epistemologies', 4. Kitchin, for example, cites Dyche 'mining Big Data reveals relationships and patterns that we didn't even know to look for', see Jill Dyche 'Big Data "Eurekas!" Don't Just Happen', *Harvard Business Review* Blog 20 (November 2012).

13 Neither is data ever simply 'given' or 'raw'; it always requires human intervention and processing. In this sense Johanna Drucker refers to 'capta' rather than 'data'. Johanna Drucker, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly* 5, no. 1 (2011): 1–21.

14 Kitchin, 'Big Data, New Epistemologies', 4. Kitchen here is referencing Donna Haraway's much earlier concept of the 'god trick': Donna Haraway, 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies* 14, no. 3 (1988): 575–99, https://doi.org/10.2307/3178066. She describes data visualisation, in particular, as 'the god trick of seeing everything from nowhere' (581).

15 See for example, Pamphlet 11 produced by the Stanford Literary Lab, which proposed a distinction between 'the published', 'the archive' and 'the corpus'. Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti and Hannah Walser, 'Canon/Archive: Large-Scale Dynamics in the Literary Field', Literary Lab Pamphlet 11 (2016), https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet11.pdf (accessed 15 June 2026).

16 Katherine Bode, 'The Equivalence of "Close" and "Distant" Reading: Or, Toward a New Object for Data-Rich Literary History'. *Modern Language Quarterly* 78, no. 1 (2017): 77–106.

17 Bode, 'The Equivalence of "Close" and "Distant" Reading', 11.

18 Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021); Nick Couldry, and Ulises A. Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (Stanford University Press, 2020); Catherine D'ignazio and Lauren F. Klein, *Data Feminism* (MIT Press, 2023).

19 Lauren Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton and David Mimno, 'Provocations from the Humanities for Generative AI Research', *arXiv* (2025), accessed 24 February 2026, https://doi.org/10.48550/arXiv.2502.19190.

20 Algee-Hewitt et al., 'Canon/Archive'.

21 Mary Poovey, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (University of Chicago Press, 1998).

22 Brake, 'Nineteenth-Century Newspaper Press Directories', 569.

23 Brake, 'Nineteenth-Century Newspaper Press Directories', 569–71, competitor publications included May's/Willing's *Press Guide* (various titles), from 1871, *Deacon's Newspaper Handbook*, 1877–1904, *Sell's Dictionary of the World Press*, 1883–1921, *Handy Newspaper List*, 1895–1915.

24 Tom O'Malley, 'Mitchell's "Newspaper Press Directory" and the Late Victorian and Early Twentieth-Century Press', *Victorian Periodicals Review* 48, no. 4 (2015): 591–606 at 593.

25 Mitchell's *Newspaper Press Directory*, 1846–47, 1851, 1854–1907 (n.p., 1846). Print. British Library Shelfmark is Mic.A.3594-3606. The example in Figure 1.1a is taken from 1908, 45 (digitised by British Library Imaging Studio; dataset with original scans available from BL Labs).

26 Transfer learning is a form of machine learning in which we first train a model on one task and adapt it for a different but related task. In the context of Natural Language Processing, this often amounts to training a model using a semi-supervised language modelling task, that is, predict the next or missing word in a sentence, after which we repurpose the language model for another task, for example named entity recognition (NER). The assumption is that information gained from one task (language modelling) will help perform another one (NER). More technically, instead of training a model from scratch each time, we transfer knowledge from a pre-trained model to improve performance on new tasks, which also requires less data or computing power.

27 Mitchell's *Newspaper Press Directory*, 1846–47, 1851, 1854–1907 (n.p., 1846). Print. British Library Shelfmark is Mic.A.3594-3606. The examples in Figure 1.3a, 1.3b and 1.3c are taken from 1908, 41.

28 https://github.com/tesseract-ocr/tesseract (accessed 15 June 2026).

29 We evaluated the performance improvement of fine-tuning Tesseract on a small sample of test data. However, because we manually corrected outcomes in a later stage – after converting all data in tabular format – we used the evaluation score primarily to select the best fine-tuning strategy, and we don't see much value in reporting them in this chapter.

30 The tabular format is, in many ways, the most useful and lasting format produced in the digitisation process. We have stored the latest and most complete version on Zenodo with DOI 10.5281/zenodo.16413810.

31 Political leaning includes: 'liberal', 'conservative', etc. Price is often notated as '1d', '2d', etc. Sometimes annual pricing is provided instead of per issue costs. Places of publication correspond with what Mitchell calls districts; these are furthermore qualified by the county they are located in.

32 Based on the cosine similarity between the vector representations of the titles.

33 Manual correction involved the following steps: we first focused on correcting the 'chain_id', which links newspaper titles across editions of the *Directories*, by sorting the resulting dataset first by 'chain_id' and then 'title' to spot linking errors and manually fix 'broken' chains. After this step, we sorted the data by the corrected 'chain_id' and moved our focus on information in other columns, such as price, place of publication and political leaning. If information is missing or diverts, we can use the context (i.e. description of the same newspaper in earlier and later years) to correct these rows. In some cases, we consulted the original scans to complete missing information or double check changes.

34 Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen and Luke Hare, 'The Past Is a Foreign Place: Improving Toponym Linking for Historical Newspapers', *Proceedings of the Computational Humanities Research Conference 2023* (2023): 368–90, at http://ceur-ws.org/Vol-3558/ (accessed 15 June 2026).

35 The *Newspaper Press Directories* are available on the British Library Shared Research Repository: https://bl.iro.bl.uk/concern/datasets/adcef12a-bb3d-40d9-871d-5784022a77e8 (accessed 15 June 2026). We also released a database containing newspapers and reference data: Kalle Westerling, Kaspar Beelen, Tim Hobson, Katherine McDonough, Nilo Pedrazzini, Daniel C.S. Wilson and Ruth Ahnert, 'LwMDB: Open Metadata for Digitised Historical Newspapers from British Library Collections', *Journal of Open Humanities Data* 11, no. 1 (2025).

36 For an overview of its shifting meaning see Angus Hawkins, '"Parliamentary Government" and Victorian Political Parties, c.1830–c.1880', *English Historical Review* 104 (1989): 638–69.

37 After normalisation of some names for example replacing '&' with 'and' and removing OCR errors.

38 In Figure 1.4 'nan' stands for not a number, i.e. occasions where no political label was detected.

39 This includes sixteen conservative to liberal switches not reported in Table 1.2.

40 Angus Hawkins, *British Party Politics, 1852–1886* (Macmillan, 1998).

41 Structural topic models discovers topics within a set of documents, while also incorporating document-level metadata (such as date, author or source). Unlike traditional topic models, a structural topic model enables researchers to analyse how topics (both their prevalence and content) vary across these different metadata variables. Structural topic modelling is particularly useful for gauging how topics change over time or differ between (political) groups, as we show in these examples. The structural topic model is currently implemented in R and more information can be found here: https://www.structuraltopicmodel.com/ (accessed 15 June 2026).

42 We stemmed the words as part of the data pre-processing pipeline.

43 This broadly confirms the classic argument of Alan J. Lee, *The Origins of the Popular Press, 1855–1914* (London: Croom Helm, 1976), 158–80.

44 See Nicolò Bonato, 'LwM Digital Residency: Visualising the Press Directories Dataset', September 2023, accessed 4 April 2025, https://livingwithmachines.ac.uk/lwm-digital-residency-visualising-the-press-directories-dataset/. By the 1920s, 'Independent' was often code for anti-Labour.

45 Beelen et al., 'Bias and Representativeness', explore these issues for the corpus of British newspapers digitised by JISC.

46 For details of the digitisation process and current reuse arrangements see, https://www.digitisednewspapers.net/histories/bl/ (accessed 4 April 2025).

47 Edmund King, 'Digitisation of Newspapers at the British Library', *Serials Librarian*, 49 (2005): 165–81, https://doi.org/10.1300/J123v49n01_07 (accessed 15 June 2026).

48 To establish the extent to which the digitised sample resembles the *Newspaper Press Directory* in its political composition, we used the Jensen-Shannon Divergence between the probability of distributions of the political labels in the digitised sample and the directories. These scores tell us to what extent the sample and population of newspaper titles resemble each other in proportional terms: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence (accessed 15 June 2026).

49 On the broader collection history, see Paul Fyfe, 'An Archaeology of Victorian Newspapers', *Victorian Periodicals Review* 49 (2016): 546–77, https://doi.org/10.1353/vpr.2016.0039.

50 Kaspar Beelen, Jon Lawrence, Katherine McDonough and Daniel C.S. Wilson, 'Whose News? Critical Methods for Assessing Bias in Large Historical Datasets', *Computational Humanities Research* 1 (2025), doi:10.1017/chr.2025.10007.

51 Andrew Hobbs, *A Fleet Street in Every Town: The Provincial Press in England, 1855–1900* (Open Book Publishers, 2018), https://doi.org/10.11647/OBP.0152.

52 Klein et al., 'Provocations from the Humanities', 9.

53 Because the BNA is not an open dataset, we have published this derived data to make our research reproducible and unlock this resource for other researchers: https://zenodo.org/records/14996278 (accessed 15 June 2026).

54 Burt L. Monroe, Michael P. Colaresi and Kevin M. Quinn, 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis* 16, no. 4 (2008): 372–403.

55 Klein et al., 'Provocations from the Humanities', 8.

56 https://github.com/Living-with-machines/subsamplr (accessed 15 June 2026).

57 https://github.com/Living-with-machines/livingwithmachines.github.io/blob/main/book/chapters/newspapers/newspaper-metadata.ipynb (accessed 15 June 2026).

58 David Woodward, ed., *Five Centuries of Map Printing* (University of Chicago Press, 1975).

59 https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~244652~5513977 (accessed on 19 June 2026).

60 Interactive versions of Figures 1.6, 1.8, 1.10, 1.12 and 1.13 are included in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

61 https://maps.nls.uk/geo/find/#zoom=9.2&lat=54.09919&lon=-2.45146&layers=101&b=10&z=0&point=54.08033,-2.80809&i=126515141 (accessed 7 April 2026).

62 For the OS, narrative histories have included W.A. Seymour, *A History of the Ordnance Survey* (Dawson, 1980), Richard Oliver, *The Ordnance Survey in the Nineteenth Century: Maps, Money, and the Growth of Government* (The Charles Close Society, 2014) and Rachel Hewitt, *Map of a Nation: A Biography of the Ordnance Survey* (Granta, 2010). The Charles Close Society has begun a carto-bibliography for the first edition of six-inch-to-one-mile scale sheets, but this is limited to only Lancashire and Yorkshire. See Rob Wheeler, 'Introduction: A Six-Inch Cartobibliography', https://www.charlesclosesociety.org/provo (accessed 18 January 2025). For a carto-bibliography of the 'old series' OS sheets (one-inch-to-one-mile), see Roger Hellyer and Richard Oliver, *The First Ordnance Survey Map: The One-Inch Old Series of England and Wales* (The Charles Close Society, 2015).

63 Both *GéoPeuple* and *GeoHistoricalData* were based at the EHESS in Paris in collaboration with IGN. See https://geopeuple.ign.fr/index.html, https://www.geohistoricaldata.org/, and http://ladehis.ehess.fr/index.php?611 (accessed 17 January 2025) for more details. On the Cassini maps, see Monique Pelletier, *Les cartes des Cassini: La science au service de l'État et des provinces*, new ed. (Éditions du CTHS, 2013).

64 https://www.davidrumsey.com/luna/servlet/s/6877hp (accessed 19 June 2026).

65 Kasra Hosseini, Daniel C. S. Wilson, Kaspar Beelen and Katherine McDonough, 'MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale' in *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '22)* (Association for Computing Machinery, 2022): 8–19, https://doi.org/10.1145/3557919.3565812.

66 Matthew Edney, *Cartography: The Ideal and Its History* (University of Chicago Press, 2019).

67 The Sanborn fire insurance map collection at the Library of Congress is an excellent example: https://www.loc.gov/collections/sanborn-maps/ (accessed 17 June 2026).

68 Chris Fleet, 'Creating, Managing, and Maximising the Potential of Large Online Georeferenced Map Layers', *e-Perimetron* 14, no. 3 (2019): 140–9; Kasra Hosseini, Katherine McDonough, Daniel van Strien, Olivia Vane and Daniel C.S. Wilson, 'Maps of a Nation? The Digitized Ordnance Survey for New Historical Research', *Journal of Victorian Culture* 26, no. 2 (April 2021): 284–99, https://doi.org/10.1093/jvcult/vcab009.

69 https://maps.nls.uk/geo/records (accessed 17 June 2026).

70 https://maps.nls.uk/geo/records/#zoom=6.0&lat=56.42669&lon=-4.95972&layers=3&b=9&point=53.8393,-2.7285 (accessed 7 April 2026).

71 Jo Guldi, 'Critical Search: A Procedure for Guide Reading in Large-Scale Textual Corpora', *Journal of Cultural Analytics* 3, no. 1 (2018): 1–34, https://doi.org/10.22148/16.030.

72 Chris Fleet, the NLS Maps Curator behind the design and development of NLS online mapping resources and collections development, worked closely with the LwM team as both facilitator and adviser.

73 https://www.maptiler.com/nls/ (accessed 17 June 2026).

74 Few libraries provide historical maps as tiles, but this approach is extremely useful in simplifying data access. More common is sharing collections as International Image Interoperability Framework (IIIF) resources: in this way, institutions provide IIIF 'manifest' URLs that allow users to access high-quality images and their metadata and associated annotations without the hassle and expense of data transfers. On the origins and design of IIIF, see Stuart Snydman, Robert Sanderson and Tom Cramer, 'The International Image Interoperability Framework (IIIF): A Community and Technology Approach for Web-Based Images', in *Archiving Conference* 12 (Society for Imaging Science and Technology, 2015): 16–21. See also https://iiif.io/. Questions remain about whether the quality of images derived from tiles or IIIF resources is sufficient for various computer vision tasks, but with further experimentation, we should have a better understanding of the opportunities and limitations. As more research engages with IIIF, we should have more data about how to work with IIIF in machine learning contexts. Tools like AIKON allow users to use IIIF resources as input data, and experiments are needed to compare how models perform on these image versions compared to 'master' images delivered via

hard drives or traditional file transfers. Ségolène Albouy, Jade Norindr, Fouad Aouinti, Clara Grometto, Robin Champenois, Alexandre Guilbaud, Stavros Lazaris, Matthieu Husson and Mathieu Aubry, *AIKON: A Computer Vision Platform for Historians* (2024), https://github.com/Aikon-platform/aikon. MapReader integrated IIIF functionality in 2025: Rosie Wood, Kasra Hosseini, Kalle Westerling, Andrew Smith, Kaspar Beelen, Daniel C.S. Wilson and Katherine McDonough, 'MapReader: Open Software for the Visual Analysis of Maps', *Journal of Open Source Software* 9, no. 101: 6434, https://doi.org/10.21105/joss.06434.

75 Chris Fleet, 'Life Beyond the Paper Graphic Index: Evaluating New Graphic Retrieval Technologies for the Future Map Library', *Liber Quarterly* 18, no. 2 (2008): 255–70, https://liberquarterly.eu/article/view/10507/11163 (accessed 2 March 2026).

76 Olivia Vane, 'Macromap: An Interactive "Small Multiples" Visualisation for Historical Map Collections' (2021), https://observablehq.com/@oliviafvane/macromap (accessed 17 June 2026). See also https://livingwithmachines.ac.uk/macromap-interactive-maps-in-time/ (accessed 17 June 2026).

77 Vane georeferenced the Old Series sheets for this work. Olivia Vane, Ordnance Survey Old/First series England and Wales 1:63360 (georeferenced sheet images) (2021), https://doi.org/10.23636/6g9b-9d72 (accessed 7 April 2026).

78 The entire collection is available at https://observablehq.com/collection/@oliviafvane/map-animations-to-publish (accessed 17 June 2026). Olivia Vane's work with NLS collections was featured as Anon. [Olivia Vane], 'How Intrepid Victorian Surveyors Mapped the Length and Breadth of Britain', *The Economist*, 6 April 2023, https://www.economist.com/interactive/britain/2023/04/06/how-intrepid-victorian-surveyors-mapped-the-length-and-breadth-of-britain.

79 On the complexity of dating OS maps, see, for example, the detailed explanation provided by Chris Fleet on the six-inch sheets: https://maps.nls.uk/os/6inch-england-and-wales/info1.html (accessed 17 June 2026).

80 https://observablehq.com/@oliviafvane/os-old-series-first-series-england-wales (accessed 19 June 2026).

81 https://observablehq.com/@oliviafvane/os-6-to-the-mile-great-britain-2nd-ed (accessed 19 June 2026).

82 Mariona Coll Ardanuy, Kaspar Beelen, Jon Lawrence, Katherine McDonough, Federico Nanni, Joshua Rhodes, Giorgia Tolfo and Daniel C.S. Wilson, 'Station to Station: Linking and Enriching Historical British Railway Data', *CHR2021: Computational Humanities Research 2021* (2021), 73, https://ceur-ws.org/Vol-2989/long_paper29.pdf. *StopsGB* stations dataset available at https://bl.iro.bl.uk/concern/datasets/0abea1b1-2a43-4422-ba84-39b354c8bb09 (accessed 17 June 2026).

83 Wood et al., 'MapReader'. On World Historical Gazetteer, see https://whgazetteer.org/ (accessed 17 June 2026).

84 Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books team and Joseph P. Pickett, 'Quantitative

Analysis of Culture Using Millions of Digitized Books', *Science* 331, no. 6014 (2011): 176–82.

85 https://living-with-machines.github.io/genre-classification/intro.html (accessed 17 June 2026).

86 https://huggingface.co/datasets/TheBritishLibrary/blbooksgenre/viewer/annotated_raw (accessed 17 June 2026).

87 These labelling functions serve more as heuristics and rough approximations. Early modern books often had very long titles.

88 Since this work was carried out there have been significant developments in the performance of instruction-tuned large language models (LLM). As a result, it would now be possible to perform this task using an LLM. However, using labelling functions can still be appealing for a number of reasons including a much lower computational cost, a higher degree of control and interpretability, and, importantly, this method allows for the incorporation of domain expertise into a machine-learning pipeline.

89 Underwood, *Distant Horizons*, 92.

90 Klein et al., 'Provocations from the Humanities', 8.

Chapter 2

MapReader: software and principles for computational map studies

Katherine McDonough, with Ruth Ahnert, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Valeria Vitale, Kalle Westerling, Daniel C.S. Wilson and Rosie Wood[1]

Introduction

Spatial data is a capacious category: it can derive from maps, catalogue data and other metadata, or fieldwork observations. Complex information about space and place is also embedded in texts, and thanks to developments in computational linguistics this data can now be unlocked at scale.[2] But as debates in the digital humanities (DH) have long recognised, scale represents both an opportunity and a challenge: a core objective of the spatial humanities has been to address the geospatial information overload faced by many fields, made all the more apparent as ever larger parts of collections in libraries, archives and other heritage sites are scanned and made available as digital collections.[3]

The Living with Machines (LwM) project was concerned with spatial data from a variety of sources, as can be seen in Chapter 1 (the press) and Chapter 3 (census returns). Here, however, we are concerned with maps as a primary source and the opportunities that big digitised map collections present for developing new computational workflows and epistemologies. The field of *computational map studies* opens up the possibility of examining the richness of the surviving cartographic archive.[4] But to do so requires a deeply interdisciplinary approach that combines the scholarly traditions of map interpretation with a new mindset for thinking *across*

maps, at the scale of entire collections, or series. In computational literary studies, questions and approaches have been scaled up to deal with massive collections of texts.[5] In the same way, computational map studies need to develop new methods appropriate to the distinct challenges of analysing large sets of maps.

This chapter introduces a tool that we developed for undertaking such work: MapReader. Our purpose here is not simply to describe this tool, and its affordances, which we have done elsewhere,[6] but also, and perhaps more importantly, to reflect on the way that implementing this methodology establishes a new *epistemological* pipeline for the historical researcher. MapReader reshapes the way that we conceptualise our research, develop our research questions, gather our evidence and build our arguments. It breaks down a method for working with maps at scale into steps that analyse abstract, but meaningful units – *patches* and *maptext* – to facilitate creative thinking about historical landscapes and cartographic practice that provides views of the past to complement non-computational analyses. As a computational method, it does not replace contextual, archival research. Bringing together both methodologies, as we do in Chapter 3, makes the most of the multiscalar inquiry – from streets to nations – that MapReader enables.

MapReader can be seen as a methodological response to two dimensions of the nineteenth-century information deluge we discuss in the book's introduction. Following its creation in 1801, at the height of the Napoleonic Wars, the British Ordnance Survey (OS) oversaw the surveying and re-surveying of Britain and Ireland at multiple scales across the nineteenth century. The first one-inch to one mile series, surveyed between 1801 and 1869, was published as ninety-one separate sheets, but later, more detailed series ran to tens of thousands of sheets. Millions of copies of these sheets were printed and sold, although it was only after the First World War that the OS began to design and market maps as an essential tool for the outdoor enthusiast.[7] Crucially for us, and thanks to the resource-intensive work undertaken by the National Library of Scotland (NLS) to scan, catalogue, georeference and 'stitch' together the georeferenced sheets of several series and editions of large-scale OS mapping, anyone can now browse these collections online, flying over the landscape like a bird. But despite the scale of the digital collections now available – over 15,000 sheets for each six-inch to one mile edition or 50,000 sheets for the twenty-five inch to one mile editions – the user is constrained to

navigate this information by searching for predetermined places of interest and looking at those maps sheet by sheet. Just as scholars of books and newspapers before us, we sought to address the challenge of how to analyse and interpret thousands of sources simultaneously using critical methods rooted in the humanities. If addressed carefully, analysing thousands of maps computationally could reframe our understanding of relationships between local, regional, national and even transnational spatial patterns. In short, cracking this challenge would enable an era of historical research where map content can motivate and address cultural, intellectual, environmental, social and political history questions that traverse spatial scales.

There is, of course, an established history of digital approaches to interpreting historical maps, but since the development of deep learning and advances in computer vision (CV), the spatial humanities have radically changed.[8] In the 2000s and 2010s, PhD students interested in using maps in coordination with digital tools were taught to use software to build Geographic Information Systems (GIS), a limited approach to structuring and organising geospatial data that was not always well suited to either historical sources or the spatial questions humanities scholars wanted to ask of them. A GIS (built with software like ArcGIS or QGIS) is at its core a spatial database. Two main data types can be stored: vector data (as points, lines and polygons) and raster data (continuous data with attributes stored for each pixel, for example, about land use classes). Visually, this data is presented as layers in the user interface, and various analyses can be used to edit, combine, transform or export selections of the data.[9] Historical maps can be loaded as images, georeferenced and then selected features of interest can be digitised manually, by tracing features like roads, for example, as lines. Scanned maps are also often simply background images, used as an alternative to a Google, Mapbox or OSM base map.

However, LwM work has been part of a wider push in the spatial humanities to reorient the field around open-source technologies, reproducible methods, and more capacious uses of maps as primary sources in digital spatial history in particular.[10] Within this wider community of practice, MapReader arose from a specific set of aims: we sought to create an iterative, reflective, collaborative, open and – in terms of its development – agile research environment for analysing a range of digitised, historical sources. We took LwM's 'radical collaboration' mindset seriously as the opportunity to work with maps from the NLS took shape.[11] This was not about historians

bringing a problem to computer or data scientists: working with maps as data was an opportunity that the historians, DH scholars and data scientists on the project co-designed. Thus, from the outset, MapReader was a unique proposition sitting between the data science community's aim to produce generalisable code, the DH penchant for experimentation, and the historian's aspiration to work outside the existing parameters defined by GIS. Furthermore, as a result of a spin-off project, Machines Reading Maps, we were also able to address a major gap in historical map data creation: maptext. Combining access to both visual and text information on maps represents a major shift towards understanding the multimodal dimensions of cartographic sources at scale.[12]

In its current form, MapReader offers historians an entry point to asking questions about groups of maps, *and* creating data needed to answer those questions. Two main tasks provide pathways to labelling visual and text information independently and together: 'patch classification' and 'text spotting'. Patches are at the heart of MapReader's approach to data. They are created by imposing a grid on top of a map sheet image – the size of the squares forming the grid can be defined by the user – and defining each grid square, or patch, as the new unit of analysis. The patch becomes the visual frame of reference that is first labelled by the researcher in the process of creating training data (annotation), and then classified automatically by the trained CV model. Labels are user defined, and developing these iteratively during the annotation stage is an essential part of the conceptual shift encouraged by MapReader. As we explain further below, labelling patches is a creative activity that every historian will (and should) approach differently. Its very ethos contrasts with the reductive approach to digitising 'what is on the map' based on (1) predetermined classes (mimicking features identified in map legends) and (2) pixel-level precision (which we contend is unnecessary for most historical analyses).

Likewise, text spotting detects and recognises text on maps ('maptext'), an element that has historically been ignored by scholars as a key aspect of map content – whether for analogue or digital analysis.[13] Text on maps focuses attention on what was labelled by cartographers: by freeing ourselves from the manual digitisation of micro-level map features, we not only accelerate the research process but also open ourselves to new opportunities to define, through an iterative process, the attributes of a map that can be meaningfully interrogated. MapReader thereby offers an epistemological space for thinking about maps as primary sources, outside

of the rules imposed by pre-packaged software from non-humanities disciplines. This opens up inquiry that is conceptually freeing (develop your own label); multiscalar (enables pattern detection across local all the way to national scales on maps like the OS large-scale sheets); and multimodal (combining visual and text data creation).

It is this epistemological shift, and the opportunities it opens for those who wish to work with maps in a digital research environment, that we focus on in this chapter. Elsewhere, we have presented the technical details for how MapReader functions and reflected on MapReader's interventions in the DH and the broader spatial disciplines.[14] MapReader offers a paradigm shift in the spatial humanities, allowing users to 'see' maps at scale in ways that adapt and refine scholarship on map understanding. MapReader is not a tool divorced from the archival reality of map collections or the topics of interest to historians across many fields, from social and political, to intellectual and environmental history, and more. If computational approaches to historical maps in non-humanities disciplines are usually geared towards the quantification of the built or natural landscape, MapReader firmly promotes a qualitative and critical view of the landscape presented on maps.

An epistemological shift in humanities software design

MapReader is part of an emerging ecosystem of 'Open Maps' tools, standards and data.[15] It represents one of the first sustained efforts to link up machine learning, CV, digitised map collections and humanities inquiry. Most previous and ongoing humanities work using CV – with the exception of library science-driven research on document processing and optical character recognition – has focused on photography, art or other audio/visual archives.[16]

Work exploring maps with CV before MapReader has largely been undertaken by two scholarly groups with different norms and aims. The first are geographic information and environmental scientists treating maps 'as data', that is, as objective spatial observations, usually to combine these with other geospatial data for research on environmental or geological resource discovery and protection, or demographic research.[17] The second community working with CV and maps have been humanities scholars. Early on this community was seeking to build Historical GIS (HGIS) of neighbourhoods, cities or regions.[18] This kind of work led in turn to the

development of 'mirror worlds' of historic cities like Amsterdam, Venice or Paris, mostly using semantic segmentation methods, which assign pixel-level labels for basic categories (such as 'land', 'building', 'water', 'road').[19] To radically oversimplify: by and large early HGIS work isolated specific features (cartographically speaking) while humanities-driven segmentation efforts have sought to understand the entire spatial footprint of the map (often across a large series).[20]

In evaluating these communities of practice on LwM, we concluded that both overlook a key lesson from the history of cartography: the ability to create digital data from map content in a way that did not suggest such content was an objective representation of a historical landscape. A *critical* approach to historical cartography – that is, critiquing the mapping process in its historical and technical contexts – is now deeply embedded in how scholars engage with maps in an analogue setting.[21] We wondered whether it would be possible to transform these practices to a digital research environment. Could there be an information processing method that was both automatic *and* critical?

A critical approach to *maps as data* was therefore the first fork in the road, setting what would become MapReader apart from other computational approaches to maps. Discussions about the complexities of working with digital facsimiles are familiar across many kinds of media, but maps face a particular challenge. If the scanned image of a map is the digital facsimile, and the physical sheet is the original, how do we refer to the landscape the map portrays? When we create datasets based on map content – commonly, but problematically, called 'extracted features' – what is the relationship of these features to the historical landscape? Where textual scholars compare facsimiles and 'originals', map users often conflate derived, 'extracted' data with the reality, or the 'ground truth', of a past landscape, skipping over the map itself as a constructed object. Extracted features then quickly calcify into facts on the ground in the absence of an approach that puts the image of the map into conversation with what is being datafied. In contrast, one of MapReader's key features is a data structure that reminds users of the distance between the map image and the new digitised data: the patch (see Figure 2.1). It sidesteps reifying derived data as a truth statement on a map by rejecting the standard forms of geospatial data, and offers the patch format as an intentionally unfamiliar transformation of scanned maps. MapReader creates a form of derived data that is not designed to help people 'use' maps as navigational

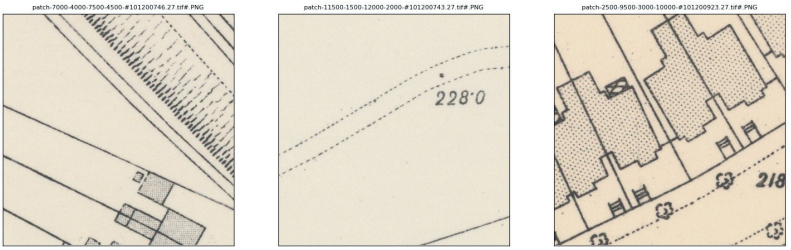


Figure 2.1: Examples of MapReader patches (with unique ID labelled above) from historical Ordnance Survey mapping. Credit: Katherine McDonough. Map image reproduced with the permission of the National Library of Scotland. CC BY-NC 4.0.

devices (as an example of one original intention for map use) but to analyse their contents as scholars.

By defining the patch as the essential data unit, MapReader takes away the impulse to delineate features and instead emphasises looking at the information contained in the patch and developing a label that is conceptually useful for a given question, and visually depicted in patches. It is represented as a raster image – that is, a digital image in which the cell values relate to the pixel colours of the original image – of a square region of a map. We turn the constraint of a defined raster space into an opportunity to engage with the map as a historical object proposing a particular vision of the landscape at a given point in time. The patch is a nod to the essential contributions of scholars of the history of cartography like Matthew Edney who have reshaped historians' understanding of mapping practice and thus maps themselves.[22] Just as the map is an argument about the landscape, so, too, is a scholar's label of a patch.

The next parting of ways from other computational approaches to maps came from a decision not to build an HGIS database and user interface. An obvious choice may have been to create a geospatial database of nineteenth-century Britain based on various layers of features identified on OS maps. Avoiding this route was, in part, a practical response to the ongoing overheads of maintenance for such outputs, and the well-documented challenges posed when the funding and expertise for such work disappears after a project ends.[23] But there was also a positive side to this, in that we wanted to open up ourselves to a more creative, experimental approach to the information that we could see on the maps, and in a way that foregrounded the analytical process. We were inspired by the Linked Open

Data approaches from Pelagios, and other projects using more flexible, *non-relational* data structures. While relational databases store data in multiple related tables, flat files store data in a single table. By choosing to store the outputs from MapReader as a flat file, we realised we could enable users to decide for themselves how they wish to relate their data (e.g. extracted features) to other historical geospatial data or knowledge. Therefore, a primary benefit of MapReader's flat data structure is that it can handle varying or evolving data schemas, making it highly flexible. It is well suited for initial data collection, qualitative analysis or situations where the structure of the data is not yet well defined or is expected to change rapidly.[24]

Building on these first two decisions, defining the 'patch' as the unit of analysis was a major shift. It introduced a form of coarse-graining that went against the highly precise data granularity of prior digital approaches to map analysis but still enabled a CV task that could infer the content on maps. But how precisely located do historians really need map content to be? What kinds of research questions lend themselves to more or less precise data requirements?

These questions had not yet been asked, despite the fact that: (1) it is more computationally expensive to create more precise data, and (2) it is not clear that such precision is useful in all cases. Previously, the assumption had been that, for cartographic documents, machine learning was simply a tool for automating whatever common tasks a GIS software user could perform. This attitude was so deep-seated that most GIScience research of the early 2000s and 2010s never considered alternatives, and the consequence was a reification and reuse of methods that create datasets that suggest locational accuracy and conform to vector data types (points, lines, polygons). Encouraging scholars to use machine learning tools as if they were merely reproducing the work that a human might do to trace features manually into a GIS is a disservice: it blurs the distinction between what a human might see and what a machine can see, but even more problematically, it prevents the human user from looking at the map with fresh eyes, and considering potential new ways of labelling map content. The scholar coming to maps as primary sources is not a 'map user': they have different needs when it comes to deciphering maps and MapReader is an attempt to offer an intellectual aid for this critical, evaluative process.

Patches offer a scalable, individualised frame through which to view historical research questions. On LwM we were interested in the evolution

of rail infrastructure and the way that it changed the built environment in Britain in the nineteenth century (which led ultimately to the work presented in Chapter 3). GIS-based approaches invite us to conceive of railways as a network of *lines*.[25] By comparison, the framework of the patch allowed us to focus on railway infrastructure embedded within the broader landscape. Mapping 'railspace' – a more holistic footprint of rail in the physical environment – is just one application of MapReader's patchwork method. For other potential users, patches represent a non-traditional way of defining the unit of analysis of map content. At scale – across thousands of maps – these two types of data usefully constrain the viewer and offer an opportunity to focus on original interpretation of what the map depicts. The patch, in particular, is a tool that by its very design pushes the user to see the map as a *historian*. Digitising map contents using traditional GIS tools steers users to adopt the tools of cartography to interpret maps, and, as Miriam Posner has put it, such software 'privilege[s] ease of use ahead of any kind of critical engagement'.[26]

MapReader is an opportunity to break out of the methodological box offered by GIS software. It opens a new future for modelling data from historical maps that responds to critiques from Posner, Drucker, d'Ignazio and Klein, and others about humanities scholars' dependency on tools from other disciplines. Making vector data easy to create, manipulate and visualise, ArcGIS and other software are what Drucker calls 'intellectual Trojan horses' because they 'conceal their epistemological biases under a guise of familiarity'. This critique matters for maps because they are already visualisations: they are already an observation from the past that the scholar seeks to analyse. In other words, when we use GIS tools to digitise content from maps as vector data, we are treating the 'observation . . . as if it were the same as the phenomena observed' and thus 'collaps[ing] the critical distance between the phenomenal world and its interpretation'. As Drucker highlights, this 'realist' approach to data is in conflict with the basic principles of the humanistic endeavour.[27]

In light of such critiques and the growing calls to enable reproducible humanities research, we made an early decision to limit MapReader's interface (and indeed, other LwM tools) to Jupyter notebooks and for all code to be open source from the start. (Jupyter notebooks are a place for storing and sharing computer code in an environment where the user can freely combine human-readable narrative with computer-readable code.[28]) But as the word 'notebook' suggests, it is a working space rather than a finished

user-oriented product. Choosing this interface meant that we could focus on the sequence of methodologies rather than the complexities of user experience. This decision also reflected our aim to position MapReader as a tool that helps humanities scholars experience CV and machine learning in action. Rather than dropping an image into a frame on a website and seeing the output, MapReader requires users to take the journey through the code to load, annotate, train/fine-tune and infer labels for patches in their set of maps. Looping through these steps not only helps them refine their knowledge of the maps' visual style, and thus also revise their questions and patch labels; it also serves a pedagogical goal of learning through repetitive action. Scholars not previously familiar with CV, machine learning or Python come away from MapReader with a significant boost to their skills. This software design approach thus responds to calls to step away from the graphical user interfaces of most GIS software and introduces what we consider to be useful friction in the MapReader pipelines. With support from documentation written by and for humanities scholars as well as data scientists, users can focus on intellectual decisions rather than technical barriers.

Because all of the code is open, it is easy for users to peek 'under the hood' and for colleagues to get involved in maintenance and development, even if it is just reporting bugs. MapReader workshops and documentation thus introduce participants to MapReader as a research tool, but also to key digital research infrastructure and concepts like versioning, code and data repositories, licensing, and open web and geospatial frameworks and standards. All of these will set them up as informed digital scholars and librarians (for example). Borrowing many best practices in open and reproducible research from *The Turing Way*,[29] we launched MapReader with an open licence, ensured versions of the code could be tracked, created mechanisms for acknowledging contributions to the project, and developed a guide for future contributors. Our commitment to openness – not only via the transparency of the method within the notebook interface but also via our open code and abundant documentation – is, unfortunately, still uncommon for research software in the humanities.[30] MapReader is thus a contribution both to computational map studies and also to DH more broadly, acting as an example of reusable research software that has expanded from its initial use case on LwM to a growing, international community of users and contributors.[31]

Process and pipeline

So, what is it like to use MapReader? How exactly do users adopt a critical, humanities mindset from the start and what does this look like throughout the entire experience? It can be useful to think about the MapReader Jupyter notebooks as the formal 'experimental' part of a research process that has preliminary and follow-up steps. Just because we are adapting to understanding maps using computational methods does not mean we are leaving behind all scholarly traditions for the study of historical cartography: but we do need to 'digitise' some of these traditions if we are working with hundreds or thousands of maps instead of only a few.

First, as explored in Chapter 1 ('The Environmental Scan as a method for digital source criticism at scale'), users should spend time getting to know their maps, their collection history and the digital assets that are (or are not) available. Ideally, while doing this, they are also establishing a working relationship with the librarian or other person curating this collection. For example, on LwM, we worked closely with Chris Fleet, the map curator at the NLS, who has shepherded through the metadata creation, digitisation and georeferencing of the NLS's OS collections. Fleet, on behalf of the NLS, was able to provide access to sheet-level metadata, imagery that was much higher quality than some other, older digitised versions of OS maps available from a national data service in the UK – as well as expertise on the maps themselves. Once we had the metadata, we were able to begin exploring the relationship between what had been scanned, what had been georeferenced and what we wanted to research. LwM team member Olivia Vane developed a suite of visualisations to enable this work.[32] They were essential to building up our knowledge of this complex digitised corpus, and informed work to digitise additional sheets that the NLS had not been able to scan. Vane's 'Macromap' visualisation, for example, uses small multiples to clearly present which OS maps in NLS collections are present and scanned.[33] For big collections, such visualisations quickly communicate to non-experts the implications of gaps in the NLS collections: as a result of documenting gaps in both the physical collection and in what had not yet been scanned (but was physically present) at the NLS, the LwM team successfully proposed a digitisation initiative at the British Library to fill gaps that would be impossible or very difficult for the NLS to address.[34] If Macromap helped define what parts

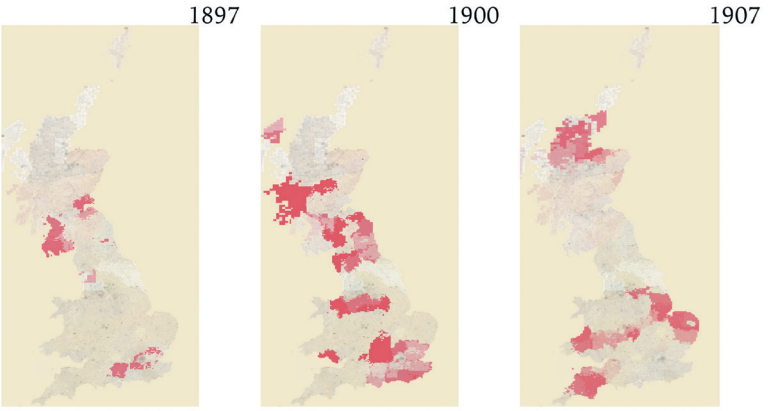


Figure 2.2: Ordnance Survey, Six-inch to one-mile maps of England, Wales and Scotland, 2nd Edition – 1887–1949. Animated chronology by publication year. The interactive version of this figure shows the six-inch sheets appearing across the map according to the year they were printed. The static version for print shows three snapshots of the publication dates for the six-inch sheets for 1897, 1900 and 1907. Credit: Olivia Vane. Map images and metadata reproduced with the permission of the National Library of Scotland. CC BY-NC 4.0.[35]

of historical OS collections were available as images, the 'Animated Map Chronologies' helped us reflect on the temporal spread of OS mapping (see Figure 2.2).

With maps and metadata in hand, we could then interrogate them through the prism of a rough research question that we proceeded to boil down iteratively to a concept, or phenomenon, associated with visual signals found on the maps.

Once these initial steps have been completed, running MapReader's patch classification pipeline usually has seven steps (see Figure 2.3).[36]

1. MapReader accepts map images as input in a few formats. On LwM, we made the decision to work with NLS sheets that were already georeferenced (e.g. the pixels of the scanned images have been transformed into geospatial coordinates).[37] But it is equally possible to work with a collection of maps that has not yet been georeferenced. If you are using georeferenced maps that are available via a tileserver (like the NLS OS maps, or

some composite maps from the David Rumsey collection) or as International Image Interoperability Framework (IIIF) resources, then you begin with the 'download' module. This process uses required fields in the metadata to define the geographical extents and zoom levels for the (physical) sheets you want to download as separate images from the 'stitched-together' tiles. Alternatively, if you are loading maps locally from your computer (or other storage), you begin with the 'load' module.[38] Loading maps and metadata is the first basic step in setting up MapReader.

2. Next, you 'patchify' your maps. This transforms your 'parent' image (one map sheet) into patches. After the selection of your maps, choosing the size of the patch is the next intellectual

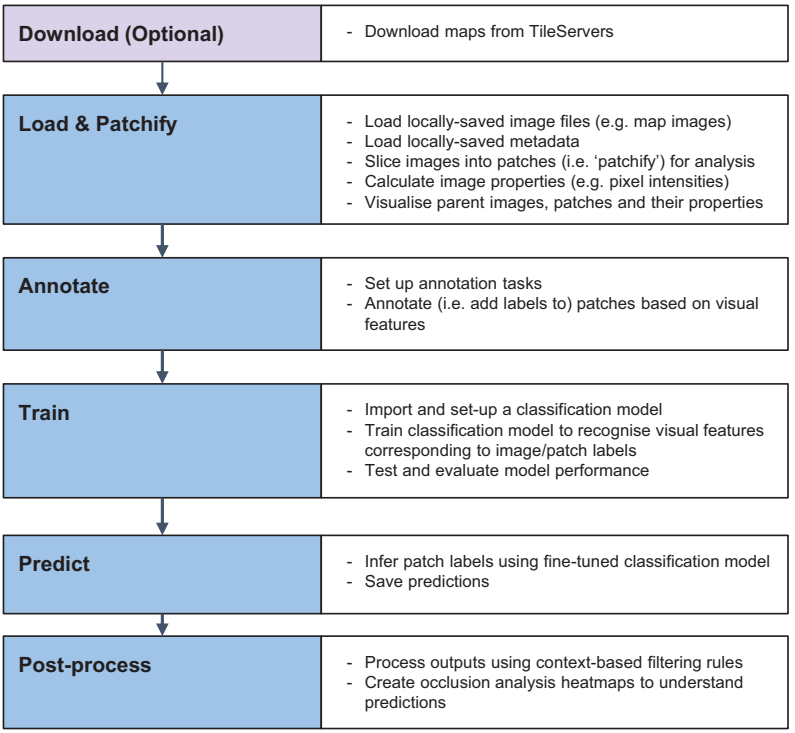


Figure 2.3: Seven MapReader steps for patch-classification pipeline. Credit: Rosie Wood. CC BY-NC 4.0.

decision you make. This is also the first place where the flexibility of MapReader comes into play: it is easy and relatively quick (depending on the size of your corpus) to 're-patchify' if you decide you want to work with a different size. (It can be handy to only experiment with a small sample of sheets initially to determine the best patch size for your research, so that changing patch sizes is quicker and more efficient.)

3. Once you have patches, you set up an annotation task. Annotating patches is the process of assigning a label to a sample of patches. In Figure 2.4, the buttons at the top contain the two user-selected labels, while the 'previous' and 'next' buttons facilitate correcting errors or skipping edge cases. The annotation interface allows users to resize the image, order patches according to a range of different metadata, and also to choose whether to include 'context' (e.g. a buffer of one extra patch around the patch in question). These labelled patches are then used for training, validation and evaluation data. Your choice of labels is the third significant intellectual choice:



Figure 2.4: Screenshot of annotation step with context turned off. Credit: Katherine McDonough. Map image reproduced with the permission of the National Library of Scotland. CC BY-NC 4.0.

mapping the concepts of interest in your research question onto the visual information on the map is not always easy. You can choose a binary yes/no-style classification task (e.g. Label 1 might be 'Concept A', and Label 2 might be 'Not Concept A'), or a multi-class task (Concept A, Concept B, Concept A & B, Not Concept A or B). You must annotate enough patches to have sufficient examples of each label to be able to divide the patches into three sets (or 'splits') for fine-tuning, testing and validating your model.

4. Next, using the training split of your annotated patches, you will fine-tune an image classification model. Poor results often mean that additional annotations are required.
5. Once you are happy with the performance of your model, you can use it to predict, or infer, the label for every patch in your set of maps. Inference results include confidence scores, so you can review low-confidence inferences to see if your model requires more examples of certain kinds of labels.
6. At the end of the prediction module, you can export your dataset as structured data (CSV, JSON). The structured data exports always contain the pixel and/or geospatial coordinates of the patch boundaries, or you can choose to export only the coordinates of the patch centroid (e.g. the geometric centre point of a patch, see Figure 2.5).

Across all these steps, a number of new files are created, both images and dataframes. If you downloaded data from a tileserver or IIIF, you will have local images and metadata for your 'parent' map sheets stored in file directories of your choice. Likewise, patch images and metadata are stored in a separate directory, each with a unique ID. The patch-level metadata file can contain key elements of the parent map metadata, such as the publication date, the URL of the original resource or the map title. Once manual annotations have been made, this creates an annotation metadata file that associates a patch's unique ID with an annotation label. Finally, after fine-tuning one or more models, you can store the details of your saved models for future use (or to share via a platform like Hugging Face, as we have). The inferred data created by the model is your final output: this is effectively the patch metadata dataframe with new columns for the inferred label and the model's confidence (out of 100 per cent).[39]



Figure 2.5: Patch centroids for railspace label in south Lancashire, near Wigan (centre left). In top centre-right, for example, the Lancashire & Yorkshire Railway Locomotive works are captured by a dense set of patches south-east of Horwich. Patches in black have the highest confidence prediction for railspace, while those in grey have lower confidence. Credit: Katherine McDonough. Map image reproduced with the permission of the National Library of Scotland. CC BY-NC 4.0.

Once you have exported your inference results, it is possible to perform any number of post-processing steps for evaluation and analysis. For example, you could calculate the surface area of your patches across the nation and compare how your concept changes over time. Or, as we have done, you can explore what text intersects with certain groups of patches.[40]

Software communities and expanding the pipeline

Beyond all the domain-specific considerations described above, MapReader was created by an interdisciplinary and experimental team

who seized the opportunity to build a piece of research software that could in fact address the needs of many scholars confronting CV for the first time, not just those exploring maps. The SciVision project at The Alan Turing Institute was the first to see the potential of applying MapReader to new imagery: in this case, inferring phenotype data in whole plant images (of *Brassica napus*). MapReader's railspace models were fine-tuned to identify branches, leaves, flowers, flower buds and pods, with robust results. Evangeline Corcoran and her co-authors highlighted MapReader's 'versatility . . . for automated analysis of images from a vast variety of scientific and humanities disciplines, but also in terms of allowing great flexibility in the modelling approaches the MapReader framework can facilitate'.[41] In short, using patches as a consistent, flexible and conceptually agnostic unit of analysis turns out to be appealing in a range of disciplinary contexts.

Since 2024, MapReader has incorporated a second task: text spotting. Text spotting started, and is still usually implemented, as a composite task, combining detection and recognition steps, using two different models. The first model detects the bounding box around a text object on the image; the second model recognises the characters of the text within the box. At the outset of LwM, we were immediately drawn to the possibility of analysing maptext, both on its own and in relation to other visual map content. We were aware of the complexities around 'OCR for maps', and experimented with existing tools like Strabo.[42] Choosing eventually to focus on the visual 'patch' classification element within LwM, McDonough won funding for the Machines Reading Maps (MRM) project (co-led with Strabo developer Yao-Yi Chiang and digital librarian Deborah Holmes-Wong) which set up the foundations for working with maptext as research data and to improve discovery in online map collections.[43] Working with the Library of Congress, the British Library and the NLS, the transatlantic MRM team not only created several new open-access datasets of maptext using the newly developed mapKurator pipeline and models, it also developed tools and methods for manually annotating maptext (with a bespoke version of the Recogito annotation tool).[44] MRM culminated with a further collaboration with the David Rumsey Historical Map Collection at Stanford Libraries, processing the text on about 57,000 georeferenced maps and making this text searchable and editable in the David Rumsey online catalogue.[45] Finally, at the conclusion of MRM, text spotting was implemented in MapReader,

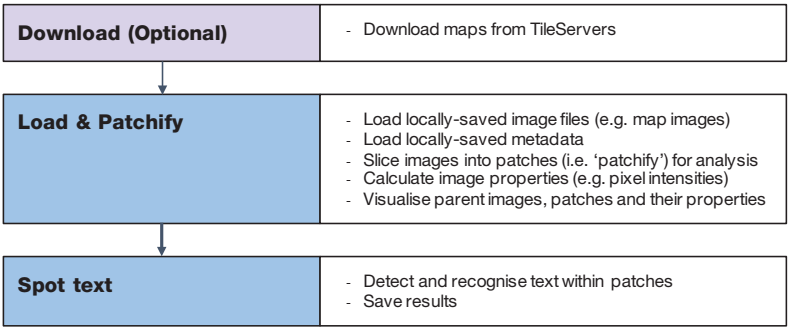


Figure 2.6: MapReader's text-spotting pipeline. Credit: Rosie Wood. CC BY-NC 4.0.

to make this more accessible to users who were not able to access the earlier MRM text-spotting tool. The text-spotting pipeline also works from patches but reconstructs words that sit across patch boundaries in a post-processing step (see Figure 2.6). With both the text-spotting and patch-classification pipelines, MapReader offers the only open-source, reproducible, multimodal approach to computational map studies.

A number of teams are using MapReader's text-spotting capabilities to find text on early modern and modern maps in Dutch, French, English and other languages.[46] Our own work is focused on examining how ordering, structuring, or otherwise organising maptext can shed light on both cartographic practice, and historical built and natural environments. Taking the next step from the 'Beyond the Tracks' chapter work to investigate the social impacts of railway infrastructure in British communities (Chapter 3), we have used text intersecting with railspace patches to differentiate between different types of railspace.[47] As one of the first studies that uses maptext not simply to explore the history of place naming, this work sets a precedent for a truly multimodal and historical exploration of map content.

MapReader emerged from the rich collaborative environment cultivated in LwM, and we continue to embrace interdisciplinary collaboration as a key to sustaining MapReader longer term. During the Data/Culture project, which followed on from LwM at the Alan Turing Institute, MapReader was featured as research software around which a programme of community building activities was designed.[48] From community calls to a series

of in-person and online workshops, this period allowed us to teach new colleagues how to use MapReader, and to learn about their interests in computational map studies. Workshops in London, as well as in Lancaster (UK), Paris, Richmond (Virginia) and at DH2025 expanded our user community and shaped recent development. Our focus in 2024–5 on implementing IIIF resources as input, for example, emerged from frequent requests from scholars and librarians for this feature, as well as the emerging ecosystem of open-source infrastructure for maps and geospatial data with IIIF at its heart. Thanks to more recent support from the Software Sustainability Institute, McDonough has been able to collaborate with the Allmaps team to link up their open-source, browser-based georeferencing tools for IIIF map collections with MapReader.[49]

Collaborating on software development is a different kind of work from the collaboration with which most humanities scholars are familiar. For the humanities-trained colleagues on the MapReader team, every stage required learning about best practices in open, reproducible software development. For our data science and software engineering colleagues, there was likewise continuous learning about library and archive collection curation and digitisation biases, and humanities scholars' ideas about meaningful concepts and research questions. A frequent refrain across LwM resonated for MapReader in particular: never assume that you and your colleague 'know' something in the same way, or that you have the same assumptions about what 'significant' claims look like. From its initial design, to its basic functionality to answer an initial research question, to maintenance considerations and training opportunities, we were always learning from each other, and the nature of this collaboration means that MapReader reflects the ambitions both of the humanities to implement methods that respect our sources and allow us to make meaningful arguments and of software engineering to build tools that can be reused and maintained easily by diverse communities.

Even as it drew on expertise from many fields, MapReader was deeply embedded in humanistic epistemologies and research processes. Laying out the path of MapReader to date so far, we have focused on how building software like MapReader is actually a humanities endeavour. But what can our approach to reading and viewing maps do in historical research? How is this different from other ways of looking at maps and why does that matter?

Using MapReader: an epistemological shift for historical research

Working with one map is not the same as working with 10,000 maps. But we do need to think about how to translate the lessons from the history of cartography to working with maps as data at scale. MapReader is just one of these ways of working with maps at scale, perhaps lighting a way through the mists of machine learning. Thinking not just about the expanse of the sheet, but about the patch as the point of departure is a fundamental shift in 'map reading'. This entails a different set of intellectual priors and outcomes, just as the computational pipeline shifts the material conditions of map reading from the maps themselves to input from a tileserver and outputs of CSV of JSON files. In so doing, it precipitates a significant epistemological shift, which we can perhaps more fully grasp if we think of it as an intellectual pipeline that corresponds to the technical pipeline that constitutes the MapReader tool.

In the twentieth century, two schools of thought emerged around map understanding. First, geographers – usually writing as self-declared cartographers, or scholars of modern cartography – had been concerned with best practices in *making* maps, which was often linked to the question of how effectively to *use* a map. Philip Muehrcke began to unpick the assumption that map-making and map-use practices were mirrors of one another when he wrote that 'map use is not the simple reverse of map making'.[50] He knew that students and colleagues needed a guide to use a map, not just make one. The second school of thought included Muehrcke's Geography colleagues at the University of Wisconsin and their students. This group are often lauded as the first historical geographers to bridge the next hurdle: from seeing maps as containers of useful, objective information, to examining them as historical documents with a variety of biases.[51] These historical geographers, writing as *scholars of maps*, developed a practice of interpreting maps not as users, but as historians.

In parallel we must think of the impact of the automation of machine cartography and the emergence of digital cartography. Following the move of command line tools to desktop applications in the 1990s, and platform-based tools in the early 2000s, mapping software like ArcGIS quickly became ubiquitous in the academic and commercial world, where it was used not simply to make maps, but principally to analyse spatial data. However, the way that this software was designed by and for cartographers

necessarily limits the possibilities of interpretation – essentially undoing the work that Muehrcke's colleagues had done to tease apart cartographic utility and *post hoc* interpretation. This is particularly strange when we consider that making maps, using maps (in everyday life or work) and understanding maps (from the past) are distinct practices. Yet scholars working on those very different processes in *digital* settings are still almost all using the same software. In other words, since the emergence of digital mapping, the capacity for doing critical humanities research using historical maps has been conceptually restricted to the ways that cartographic design wants users to engage with maps.

MapReader is designed especially to address the need to *understand* maps from a humanistic perspective. It develops an approach to step outside the bounds of the data structures and tools that link map design and spatial analysis, offering an alternative to the software that was designed to make maps and analyse contemporary spatial data. MapReader enables users to regard the map as an argument, or a proposition, as we have become accustomed to do in non-computational map interpretation, and to avoid the pitfalls of computational map processing that treat the map as a mirror of the landscape.[52] For MapReader, patches and text are the alternatives to seeing 'features' as the components that make up a map.[53] Instead of expecting to find discrete features, patches open up a space (literally) to reconceptualise what we see on the map, and what we call it. Expecting objectivity from a map suggests that there are objects contained within that can be found and extracted: this might be true for the map's original users (though they, too, should be attuned to a map's biases), but reading a map as its creators intended is not the only way to approach it as a source. Furthermore, by using maps in unintended ways, we look and see not as cartographers, but as historians reading against the grain. By changing our expectations and asserting a critical stance vis-à-vis the map, we no longer assume that a map can be teased apart into wholly distinct objects or features, nor that what we are doing is removing data from its context (as suggested by the dominant language of data 'extraction').[54]

Working with patches as a unit of analysis changes the way one asks and refines research questions, reconfigures what phenomena are of interest on a map, and shapes interpretation. To understand how this works, we might compare the epistemological process to discussions in computational literary studies around 'algorithmic criticism'. After coining this term, Stephen Ramsay wrote that 'if text analysis is to participate in literary

critical endeavor in some manner beyond fact-checking, it must endeavor to assist the critic in the unfolding of interpretative possibilities'.[55] Replace 'text' with 'map', 'literary critical' with 'historical', and 'critic' with 'historian', and we have a good basis for what MapReader does with patches. MapReader enables *reading* of patches, and therefore interpretation, because we are not pre-defining the object that is present (on the map, but also, by extension, in the historical world), nor are we assuming that a patch can only be classed in one way. This 'deformation' of the map is what sets MapReader apart from other map data processing pipelines that extract vector data or segment sheets into standardised classes. A patch is an open space, and its label is whatever suits the scholar, as long as there is a relationship between that concept and the visual information within the patch. There are as many ways to read a map as there are questions about maps. And, as Ramsay argued, it is the 'rigid, inexorable, uncompromising logic of algorithmic transformation as the constraint under which critical vision may flourish'.[56]

Furthermore, for MapReader, modelling maps as sets of patches is a way to avoid the error of assuming that the map represents the historical landscape. Mark Algee-Hewitt makes a similar pitch for literary studies when he suggests that we must 'let go of the idea that we know our artifacts well'.[57] Instead, as he argues, 'by complicating our understanding of a phenomenon like authorship through a computational model, we give ourselves the opportunity to revise our understanding of the underlying phenomena outside the constraints of habituation and practice such that we can radically alter our understanding of the object itself'.[58] In terms of maps, the opportunity for revision comes with letting go of the idea of discrete map features as the model for the digital artefact, or indeed as a digital placeholder for something 'real'. With MapReader the transformation of the map into patches creates an unexpected object, one that we are not trained to analyse.

The positive consequence of this is that it reconfigures what it means to *look* at a map. In particular, MapReader addresses what Katherine Bode has called for as methodological interventions that are 'simultaneously technological and theoretical': MapReader is designed around the fact that scholars are interacting with digital objects with a host of attendant digital infrastructures remediating scholarly interactions with these collections.[59]

This new shape, or object of scrutiny, will not meet the needs of every historian, every map or every question. MapReader should be viewed as

a critical intervention offering an alternative to the assumptions inherent to the design of GIS software and industry-standard formats, but, it can also operate as a complement to these where appropriate. Our work in Chapter 3 is a case in point: much of the downstream analysis of the 'converged' railspace and street-level census data used QGIS as well as Pythonic geospatial libraries like geopandas. MapReader offers an experimental space in which to test, in an iterative and visual-first environment, just how much precision one needs to address the issues of interest. This is possible through choosing the size of the patch (in pixels or metres), as well as the labels.

We can explore these principles and practices through the example of LwM's railspace concept. In previous digital explorations of the history of rail, the focus has been on mapping and reusing vector data for railway tracks, often to answer socio-economic history questions.[60] In contrast, our question was about the lived experience of rail at the height of its activity in Britain. Historians had been aware, through neighbourhood-, town- or city-based case studies, that the arrival of rail had increased class-based social segregation in British communities. But was this a uniform process nationally, across urban, suburban and rural areas? And what of the distinction between living near a station versus living in close proximity to major railroad works? In short, we sought to investigate what it was like to live near (or far) from rail both as a mode of transport (e.g. measured by distance from a passenger station) and also as a major infrastructural presence throughout the landscape. We arrived at this question as a group of social and cultural historians interested in the interplay between politics, the built environment and technology after intensively studying what OS maps were accessible and what seemed like a timely opportunity to connect historical microcensus data to questions about residential class segregation. MapReader's patches reinforce that we are not extracting tracks from maps as vector data, but locating the footprint of railway infrastructure among its surroundings, which we called *railspace*.

Railspace is only visible *as* patches, not in line data. The concept emerged from the constraints of this digital transformation of the map sheet: and thus our research question was refined as we adapted to working with patches. Unlike vector data, these 100 metre by 100 metre frames around segments of the map retain a connection to the map image, and repeatedly remind the viewer that rail infrastructure sits within a complex built and natural environment. Railspace becomes the set of patches that

a classification model can identify as sharing certain signals, as identified during the annotation of training data. This training data was generated by the historians on the project annotating patches in which they observed features they identified as 'rail', but the way that the model works – by 'noticing' patterns across large-scale data – means that it learns not only how rail is represented visually on maps but also the wider contexts in which it habitually appears. The method thereby returns attention to context in two senses, both on the part of the model and on the part of the scholar, who is encouraged by the nature of the annotation task to scan the entire patch, taking in that wider framing. As such, this highly computational pipeline in fact enables a way of looking at maps that is more similar to the ways historians traditionally observed maps, before GIS reduced objects to lines and points.

Far from being an unthinking mechanical activity, therefore, the patch annotation task is essential to the intellectual process of concept definition. This is supported by the iterative nature of the annotation step in the pipeline. By this we mean that it is possible, indeed encouraged, to use this step of the pipeline to experiment with the name of your label, the guidelines you set for what the visual requirements are for a given label, whether or not your set of labels is sufficient for your needs and, ultimately, whether your research questions also need revising. For scholars still getting to know their maps well, annotation offers a chance to look, patch by patch, across a corpus. In this way, annotation eventually consolidates the scholar's conceptualisation of their labels, while also identifying edge cases (the patches that one struggles to label definitively), which may be historically significant. For LwM, annotating railspace patches first collectively and then individually highlighted another affordance of this approach: annotations with a team can be a productive way to debate conceptual definitions, while one person's annotations reflect their unique interpretation of labels. Options for labelling practices can thus introduce opportunities for establishing shared understandings of novel concepts, or enable wholly personal 'views'.[61] This is a major contribution: MapReader enables fast, reproducible, bespoke classifications of map content through patches, facilitating historical arguments that may be (as historical claims often are) attached to one scholar, or to many.

But how does the output of the pipeline correspond to what a historian might regard as an outcome? MapReader patches whose labels are inferred by a model are, like their parent maps, a proposition based on

the labelling decisions of the user. In other words, the outputs are not the answer. Just like the notes a historian might take about the documents they represent a reading; MapReader's outputs constitute an interpretative layer, or 'screen' superimposed on the map.[62] They also resemble a historian's notes in the way they move the scholar one step closer to a domain-specific claim. The inferred label orients the scholar's attention towards that claim, but the distance between the (potentially) millions of labelled patches and an argument that is recognisable to historians, digital or not, remains to be covered.[63] For our railspace work, we refined our understanding of the presence of 30.5 million railspace patches by calculating their proximity to hundreds of thousands of residential streets in England, Wales and Scotland: this gave us something to measure that linked people to the spatial footprint of rail infrastructure. From there, we could analyse different groups based on railspace proximity, census attributes and prior knowledge from earlier historiography. Crafting this bridge from nationwide patch data to fictional and real-life examples of the experience of living on specific streets heavily impacted by railspace thus combined our domain expertise with the model outputs.

Because they are inferences, these automatically labelled patches are imperfect. Machines make mistakes, but these are based on patterns in the data as seen by the model.[64] While our railspace experiments had extremely strong performance when evaluated against expert-annotated patches (between 95 and 99 per cent F1 scores across the different labels) some errors will nevertheless persist.[65] Another way that MapReader is an iterative process is that often model fine-tuning will surface systematic false positives or false negatives. It was essential to view outputs and scan the regions for patterns in these errors. For railspace, false positives included hatched rock outcroppings along coasts and in the Highlands, as well as drainage ditches in East Anglia. False negatives were usually in hyper-urban areas with lots of other infrastructure, urban trams (which we had decided to exclude from our railspace category) and rail tunnels (which we accepted as non-railspace in the end). These discoveries necessitate further annotation, both random and targeted. On the other hand, these mistakes can be fruitful, pointing to cartographic practices that the user might not yet have noticed, or to underappreciated variations in the concept of interest.

Could we have made the same arguments about residential segregation without the railspace dataset? No: first, railspace emerged from

our *engagement* with MapReader; second, there is no other way we could simultaneously explore the hyper-local, street-level, city, county and national data. The continuity of OS map data is unique in providing extremely close-up views of a rapidly changing landscape, but we would never have been able to undertake a (multi)national-scale project without the speed and efficiency of MapReader. Where the historian chooses to take things from here is a separate matter. As Chapter 3 shows, working with nationwide map data surveyed at a large scale allows us to observe spatial patterns at multiple levels. We can make macro-level arguments (e.g. observations at the level of the nation about how people of different socio-economic classes were situated in relation to the amenities and disamenities of rail), meso-level arguments (e.g. observing regional trends or deviations) as well as using inferences to direct us to localised case studies and further deep archival work. As such the distant view enabled by MapReader is highly supportive of close reading and the traditional archival practices of historical research.

Paths forward

While MapReader has been around now for a few years, in many ways it still feels like the beginning. Wood has been working hard to make the code sustainable; Westerling ramped up efforts on a range of documentation; Wood and Lilley have been implementing MapReader in a range of large-scale computing environments in the UK;[66] and Wilson, Beelen, and McDonough have been learning to work with maptext as research data, and to compare change over time across OS editions.[67] The workshops and community calls have cultivated a broad, interdisciplinary group of users and even some early contributors to our code and documentation, such as David Alexander from the UK Peak District National Park Authority.[68] Collaborations with peer open-source software for browser-based georeferencing (Allmaps) and manual annotation (in collaboration with Rainer Simon), partnerships with a range of map libraries, and emerging digital research infrastructure support for humanities software in the UK are all shaping MapReader's future roadmap.[69] All this goes to show that building sustainable software is a community endeavour.

Historical maps are important, not just for historians and geographers but for anyone who seeks to understand the built and natural

environments of the past. MapReader helps people use these digitised collections responsibly and creatively. While we have showcased the LwM work with OS maps here, projects around the world are using MapReader on other series maps (such as the Sanborn fire insurance maps in the US) and non-serial collections that are much more heterogeneous in style (for example, the GLOBALISE project's analysis of all Dutch East India Company maps).[70] MapReader remains extremely well suited to large-scale topographical or urban maps which were printed as sets of hundreds or thousands of sheets representing continuous or quasi-continuous land. But others have experimented with MapReader on very small collections of maps, making use of the pipeline only up to the point of annotation, and simply annotating *all* patches in their dataset. This additional flexibility means that MapReader is not only an AI tool: it can be adapted for wholly manual annotation tasks where patches are a useful framework. These capabilities will open doors for historical map interpretation across the disciplines.

We believe that this is an important contribution to the wider developments starting to emerge in computational map studies. The Venice Time Machine team working with Venetian and Swiss maps, the Mapping Inequality team's work with text on Sanborn maps, and the organisers of the MapText competition (as part of the International Conference on Document Analysis and Recognition's [ICDAR] annual set of competitions to benchmark and improve models for document processing tasks) are all making contributions to methods and data and are enhancing the role of maps in historical and cross-disciplinary research.[71]

Thinking more broadly about the epistemological underpinnings of MapReader, it would be shortsighted to ignore the further possibilities to use its pipeline with non-historical map imagery. Within the MapReader team Lilley has experimented with patchifying and annotating LiDAR imagery (laser technology that generates 3D models of the Earth's surface), and we are now exploring using a range of other remote sensing data for further comparison of change over time in environmental features such as likely historical hedgerow locations in England. As we learned from the collaboration with SciVision, MapReader applications are not even always geospatial, and this openness to experimental applications in biology, for example, has helped us to better understand what models 'see', and how there can be a two-way dialogue between the natural sciences and the humanities.

MapReader began life as an experimental opportunity on LwM, and this mindset permeates our ongoing ethos after the end of that project. Directed by McDonough, the team's aspirations are not just to provide open-source, reproducible code from our own LwM research, but to demonstrate and train others to shift how they think about what we can do with big image data in the humanities. Embroiled in a host of ethical, environmental and intellectual debates, AI writ large is a daunting challenge of our age. MapReader applies non-generative classification, detection and recognition models, and we have intentionally limited MapReader to these tasks for the time being. We could not do this work without the assistance of machine-learning-driven inference, but for both environmental and intellectual reasons we resist the temptation to apply the latest state-of-the-art generative AI methods. MapReader's patch classification task, in particular, may be positively naïve from a CV perspective, but this simplicity is its virtue. By relying only on the human annotation of patches for training, rather than generative AI, we are able to understand, refine and test our classifications – and thereby begin to open the black box. Algee-Hewitt has written of the computational humanities as a chance to 'use the methods and strategies of computation to put pressure on our received theoretical and historical objects and [to] uncover alternative ways of thinking'.[72] Patch classification does just this. Detecting maptext also (perhaps unexpectedly) uncovers an entire world of largely ignored cartographic detail. Both tasks require inference to be applied at scale. In doing so we are indeed using AI to see our sources anew, but our non-generative AI approach centres historical expertise in data creation and its interpretation.

To accomplish all of this, we depended on the radical collaboration at the heart of LwM.[73] Building MapReader, using it for novel research, and teaching others how to use it has been an exercise in reconceiving what it means to be a historian, a digital humanist or a data scientist. Writing GitHub tickets, drafting and reviewing code, annotating patches, emailing librarians, meeting on Zoom to debate label guidelines, more meeting on Zoom to talk about development priorities, co-writing articles: there is no better evidence than these shared activities of just how much MapReader is the product of the intensive exchange of ideas over years. It may just be a pile of code, JSON files and geotiffs, but MapReader has shaped our scholarly and professional trajectories for the better, and we hope it will continue to have a positive impact in DH, history and beyond for many years to come.

Notes

1 McDonough planned and wrote the chapter. Ahnert collaborated on chapter scoping, writing and editing; Lawrence contributed to editing; and all authors contributed to MapReader's conceptualisation, shared reflections used throughout the chapter and reviewed the final version. After the first author, names are alphabetical.

2 Key early examples of work applying these methods include Joanna E. Taylor and Ian N. Gregory, *Deep Mapping the Literary Lake District: A Geographical Text Analysis* (Bucknell University Press, 2022), https://doi.org/10.2307/j.ctv2v55bsf; Patricia Murrieta-Flores and Ian N. Gregory, 'Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology', *Open Archaeology* 1, no. 1 (2015): 166–75, https://doi.org/10.1515/opar-2015-0010; Jim Clifford, Beatrice Alex, Colin M. Coates, Ewan Klein and Andrew Watson, 'Geoparsing History: Locating Commodities in Ten Million Pages of Nineteenth-Century Sources', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49, no. 3 (2016): 115–31, https://doi.org/10.1080/01615440.2015.1116419; Catherine Porter, Paul Atkinson and Ian N. Gregory, 'Space and Time in 100 Million Words: Health and Disease in a Nineteenth-Century Newspaper', *International Journal of Humanities and Arts Computing* 12, no. 2 (2018): 196–216, https://doi.org/10.3366/ijhac.2018.0222.

3 Like scale, the lack of methods, infrastructure, underlying data and models for non-English and low-resource languages also impedes applications in many fields. Earlier work by members of this team sought to extend geoparsing (largely only applied to English text corpora) to early modern French texts. See Katherine McDonough, Ludovic Moncla and Matje van de Camp, 'Named Entity Recognition Goes to Old Regime France', *International Journal of Geographical Information Science* 33, no. 12 (2019): 2498–522, https://doi.org/10.1080/13658816.2019.1620235.

4 Katherine McDonough and Valeria Vitale, 'Introduction', in 'Maps and Machines: Recent Perspectives on Humanities Research Using AI and Maps', ed. Katherine McDonough and Valeria Vitale, Forum, *Imago Mundi* 76, no. 2 (2024): 282, https://doi.org/10.1080/03085694.2025.2453329.

5 Scholars have addressed this in terms of text of course, but also for the page as an image. For the latter, see James Dobson and Scott Sanders, 'Distant Approaches to the Printed Page', *Digital Studies / Le Champ Numérique* 12, no. 1 (2022), https://doi.org/10.16995/dscn.8107; Andrew Piper, Chad Wellmon and Mohamed Cheriet, 'The Page Image: Towards a Visual History of Digital Documents', *Book History* 23, no. 1 (2020): 365–97, https://doi.org/10.1353/bh.2020.0010.

6 Kasra Hosseini, Katherine McDonough, Daniel van Strien, Olivia Vane and Daniel C.S. Wilson, 'Maps of a Nation? The Digitized Ordnance Survey for New Historical Research', *Journal of Victorian Culture* 26, no. 2 (2021): 284–99, https://doi.org/10.1093/jvcult/vcab009; Kasra Hosseini, Daniel C.S. Wilson, Kaspar Beelen and Katherine McDonough, 'MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale', *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities* (2022), 8–19,

https://doi.org/10.1145/3557919.3565812; Rosie Wood, Kasra Hosseini, Kalle Westerling, Andrew Smith, Kaspar Beelen, Daniel C.S. Wilson and Katherine McDonough, 'MapReader: Open Software for the Visual Analysis of Maps', *Journal of Open Source Software* 9, no. 101 (2024): 6434, https://doi.org/10.21105/joss.06434.

7 Richard Oliver, *The Ordnance Survey in the Nineteenth Century: Maps, Money, and the Growth of Government* (The Charles Close Society, 2014), 505.

8 For earlier reviews of digital approaches to interpreting historical maps, see Anne Kelly Knowles, 'Emerging Trends in Historical GIS', *Historical Geography* 33 (2005): 7–13.

9 For classic early examples of using GIS for historical research (published by the company that sells ArcGIS), see Anne Kelly Knowles, ed., *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (ESRI, 2008).

10 See, for example, Valeria Vitale, Pau de Soto, Rainer Simon, Elton Barker, Leif Isaksen and Rebecca Kahn, 'Pelagios – Connecting Histories of Place. Part I: Methods and Tools', *International Journal of Humanities and Arts Computing* 15, nos. 1–2 (2021): 5–32, https://doi.org/10.3366/ijhac.2021.0260; Rombert Stapel and Ivo Zandhuis, 'Linked Data for Modelling and Replicating the Knowledge Production Process in Data-Driven Humanities Research', *Digital Scholarship in the Humanities* 40, supp. 1 (2025): i100–i107, https://doi.org/10.1093/llc/fqae038; and Niamh NicGhabhann Coleman, Zenobie Garrett and Frances Kane, '"That's a Powerful Map": Shared Authority, Public Engagement, and the Archives of the First Ordnance Survey of Ireland', *Journal of Historical Geography* 88 (June 2025): 65–73, https://doi.org/10.1016/j.jhg.2024.12.006.

11 For further discussion of the project's collaborative practice, see Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009175548.

12 On the multimodal shift in DH see Thomas Smits and Melvin Wevers, 'A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections', *Digital Scholarship in the Humanities* 38, no. 3 (2023): 1267–80, https://doi.org/10.1093/llc/fqad008; and Taylor Arnold and Lauren Tilton, *Distant Viewing* (MIT Press, 2023).

13 Katherine McDonough, Kaspar Beelen, Daniel C.S. Wilson and Rosie Wood, 'Reading Maps at a Distance: Texts on Maps as New Historical Data', *Imago Mundi* 76, no. 2 (2024): 296–307, https://doi.org/10.1080/03085694.2024.2453336.

14 Katherine McDonough, 'Maps as Data', in *Computational Humanities*, ed. Lauren Tilton, David Mimno and Jessica Marie Johnson (Minnesota University Press, 2024), 99–126, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/992f5fc5-9a0d-4e9a-863a-adc1de9a678e.

15 *Open Maps Meeting*, accessed 18 February 2026, https://openmapsmeeting .nl/.

16 See, for example, Matthew Lincoln, Julia Corrin, Emily Davis and Scott B. Weingart, 'CAMPI: Computer-Aided Metadata Generation for Photo Archives Initiative' (2020), https://doi.org/10.1184/R1/12791807; Taylor Arnold and Lauren Tilton, 'Distant Viewing: Analyzing Large Visual Corpora', *Digital Scholarship in the Humanities* 34, supp. 1 (2019): i3–16, https://doi.org/10.1093/llc/fqz013. More recently: Leonardo Impett and Fabian Offert, 'There Is a Digital Art History', *Visual Resources* 38, no. 2 (2022): 186–209, https://doi.org/10.1080/01973762.2024.2362466; Arnold and Tilton, *Distant Viewing*. As an exception, see the wide-ranging studies in Kevin Kee and Timothy Compeau, eds, *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History* (University of Michigan Press, 2019), https://doi.org/10.3998/mpub.9964786.

17 Yao-Yi Chiang and Craig A. Knoblock, 'Recognizing Text in Raster Maps', *Geoinformatica* 19, no. 1 (2015): 1–27, https://doi.org/10.1007/s10707-014-0203-9; Johannes H. Uhl, Stefan Leyk, Yao-Yi Chiang and Craig A. Knoblock, 'Towards the Automated Large-Scale Reconstruction of Past Road Networks from Historical Maps', *Computers, Environment and Urban Systems* 94 (2022): 101794, https://doi.org/10.1016/j.compenvurbsys.2022.101794; Johannes H. Uhl, Stefan Leyk, Yao-Yi Chiang, Weiwei Duan and Craig A. Knoblock, 'Automated Extraction of Human Settlement Patterns from Historical Topographic Map Series Using Weakly Supervised Convolutional Neural Networks', *IEEE Access* 8 (2019): 6978–96, https://doi.org/10.1109/ACCESS.2019.2963213; Yao-Yi Chiang, Weiwei Duan, Stefan Leyk, Johannes H. Uhl and Craig A. Knoblock, *Using Historical Maps in Scientific Studies: Applications, Challenges, and Best Practices* (Springer, 2020), https://doi.org/10.1007/978-3-319-66908-3.

18 On HGIS scholarship in general see Jordi Martí-Henneberg, 'Geographical Information Systems and the Study of History', *The Journal of Interdisciplinary History* 42, no. 1 (2011): 1–13, https://doi.org/10.1162/JINH_a_00202; Anne Kelly Knowles, 'Historical Geographic Information Systems and Social Science History', *Social Science History* 40, no. 4 (2016): 741–50, https://doi.org/10.1017/ssh.2016.29.

19 Isabella di Lenardo and Frédéric Kaplan, 'Venice Time Machine: Recreating the Density of the Past', paper presented at the annual ADHO Digital Humanities Conference, Sydney, 2015; Frédéric Kaplan and Isabelle de Lenardo, 'Building a Mirror World for Venice', in *The Aura in the Age of Digital Materiality: Rethinking Preservation in the Shadow of an Uncertain Future* (Factum Foundation, 2020): 197–201; Frédéric Kaplan and Isabella di Lenardo, 'The Advent of the 4D Mirror World', *Urban Planning* 5, no. 2 (2020): 307–10, https://doi.org/10.17645/up.v5i2.3133; Beatrice Vaienti, Rémi Petitpierre, Isabelle di Lenardo and Frédéric Kaplan, 'Machine-Learning-Enhanced Procedural Modeling for 4D Historical Cities Reconstruction', *Remote Sensing* 15, no. 13 (2023): 3352, https://doi.org/10.3390/rs15133352. For an additional, recent, approach for Paris, see the ANR project *SoDuCo ('Social Dynamics in Urban Context: Open Tools, Models, and Data – Paris and Its Suburbs, 1789–1950')*, accessed 18 February 2026, https://soduco.geohistoricaldata.org/. SoDuCo publishes their datasets openly at https://

soduco.geohistoricaldata.org/datasets/. For example, *Annuaires Historiques Parisiens, 1798–1914. Extraction Structurée et Géolocalisée à l'Adresse des Listes Nominatives par Ordre Alphabétique et par Activité dans les Volumes Numérisés*, 2023, accessed 18 February 2026, https://doi.org/10.34847/NKL.98EEM49T. See also SoDuCo GitHub Repository, accessed 18 February 2026, https://github.com/soduco.

20 Sofia Ares Oliveira, Isabelle di Lenardo, Bastien Tourenc and Frédéric Kaplan, 'A Deep Learning Approach to Cadastral Computing', paper presented at the annual ADHO Digital Humanities Conference, Utrecht, 2019; Rémi Petitpierre, Frédéric Kaplan and Isabella di Lenardo, 'Generic Semantic Segmentation of Historical Maps', *CHR 2021: Computational Humanities Research Conference Proceedings* (17 November 2021): 228–48, https://ceur-ws.org/Vol-2989/long_paper27.pdf; Rémi Petitpierre and Paul Guhennec, 'Effective Annotation for the Automatic Vectorization of Cadastral Maps', *Digital Scholarship in the Humanities* 38, no. 3 (2023): 1227–37, https://doi.org/10.1093/llc/fqad006.

21 Matthew H. Edney, 'Academic Cartography, Internal Map History, and the Critical Study of Mapping Processes', *Imago Mundi* 66, supp. 1 (2014): 83–106, https://doi.org/10.1080/03085694.2014.947853; Matthew H. Edney, 'JB Harley (1932–1991): Questioning Maps, Questioning Cartography, Questioning Cartographers', *Cartography and Geographic Information Systems* 19, no. 3 (1992): 175–8, https://doi.org/10.1559/152304092783762290.

22 Matthew H. Edney, *Cartography: The Ideal and Its History* (University of Chicago Press, 2019); see also, in general, J.B. Harley's work, collected in *The New Nature of Maps: Essays in the History of Cartography* (Johns Hopkins University Press, 2002).

23 On these challenges and possible solutions, see *The Endings Project: Building Sustainable Digital Humanities Projects*, accessed 18 February 2026, https://endings.uvic.ca/index.html.

24 This approach also means that MapReader outputs can be processed elsewhere and then reused in minimal computing environments where working with very large image data or hosting outputs in a database would not be possible. The scale of the output files remains an issue, however, and large data transfers and tools for visualising and downstream analysis of the data could still require significant compute and internet resources. On minimal computing see the *DHQ* special issue on this topic, starting with Roopika Risam and Alex Gil, 'Introduction: The Questions of Minimal Computing', *Digital Humanities Quarterly* 16, no. 2 (2022), https://dhq.digitalhumanities.org/vol/16/2/000646/000646.html.

25 The CAMPOP project has deposited static snapshots of railway tracks as vector data for 1851, 1861 and 1881 on the UK Data Service. See Chapter 3, note 42, for details. For a reference to this data and its origins, see Dan Bogart, Xuesheng You, Eduard J. Alvarez-Palau, Max Satchell and Leigh Shaw-Taylor, 'Railways, Divergence, and Structural Change in 19th-Century England and Wales', *Journal of Urban Economics* 128 (2022): 103390, https://doi.org/10.1016/j.jue.2021.103390.

26 Miriam Posner, 'What's Next: The Radical, Unrealized Potential of Digital Humanities', in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/a22aca14-0eb0-4cc6-a622-6fee9428a357. See a longer discussion of this issue in McDonough, 'Maps as Data'.

27 Johanna Drucker, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly* 5, no. 1 (2011), https://dhq-static.digitalhumanities.org/pdf/000091.pdf. We do not adopt Drucker's language of *capta*, but the spirit of this intervention and the critical thinking about the nature of data is embedded in the design and use of MapReader. For further reflection on data versus capta, see Matthew Lavin, 'Why Digital Humanists Should Emphasize Situated Data over Capta', in *The Companion to Digital Humanities in Practice*, ed. Constance Crompton, Laura Estill, Richard J. Lane and Ray Siemens (Routledge, 2025), 121–36, https://doi.org/10.4324/9781003327677.

28 For an accessible introduction to Jupyter notebooks, see Quinn Dombrowski, Tassie Gniady and David Kloster, 'Introduction to Jupyter Notebooks', *The Programming Historian* (2019; updated 2025), https://doi.org/10.46430/phen0087.

29 The Turing Way Community, *The Turing Way: A Handbook for Reproducible, Ethical and Collaborative research*, Version 1.2.3 (2025), https://doi.org/10.5281/zenodo.15213042.

30 Zotero, Tropy and Omeka and other (now) Digital Scholar-based software (https://digitalscholar.org/), for example, are key examples of open-source humanities software. Zotero, v. 9.0.3, Corporation for Digital Scholarship, released 2026. We also looked to work like Johanna Drucker, 'Sustainability and Complexity: Knowledge and Authority in the Digital Humanities', *Digital Scholarship in the Humanities* 36, supp. 2 (2021): ii86–94, https://doi.org/10.1093/llc/fqab025; Giles Bergel, Pip Willcox, Guyda Armstrong, James Baker, Arianna Ciula, Nicholas Cole, Julianne Nyhan, et al. 'Sustaining Digital Humanities in the UK', Zenodo, 25 September 2020, https://doi.org/10.5281/zenodo.4046267.

31 *MapReader Contributors' Code of Conduct*, accessed 18 February 2026, https://github.com/maps-as-data/MapReader/blob/main/CODE_OF_CONDUCT.md. *MapReader Contributors*, accessed 18 February 2026, https://github.com/maps-as-data/MapReader/blob/main/contributors.md.

32 These interactive notebooks were a crucial part of our thinking process for understanding the contours of the collection and the complex temporal dynamics of these. The certainty that these will, over time, become dead links has been addressed by the use of static figures in the print copy of the book. Future accessibility is also addressed by sharing the underlying code and files used to create these visualisations as web archive links. Olivia Vane, 'Animated Map Chronologies' [collection], last modified 11 April 2023, accessed 23 February 2026, https://observablehq.com/@oliviafvane/animated-map-chronologies?collection=@oliviafvane/map-animations-to-publish, and specifically the 'OS 6"-to-the-mile, Great Britain: 2nd Ed' notebook, accessed 23 February 2026, https://observablehq.com/@oliviafvane/os-6-to-the-mile-great-britain-2nd-ed?collection=@oliviafvane/map-animations-to-publish.

Alternative archived version of the files on this page: https://web.archive.org/web/20260219100153/https://api.observablehq.com/@oliviafvane/os-6-to-the-mile-great-britain-2nd-ed@280.tgz?v=3.

33 Olivia Vane, 'Macromap: An Interactive "Small Multiples" Visualisation for Historical Map Collections', last modified 21 November 2021, https://observablehq.com/@oliviafvane/macromap. Alternative archived version of the files on this page: https://web.archive.org/web/20260223122942/https://api.observablehq.com/@oliviafvane/macromap@7401.tgz?v=3. See also Olivia Vane, 'Macromap: Interactive Maps in Time', last modified 16 November 2021, accessed 23 February 2026, https://livingwithmachines.ac.uk/macromap-interactive-maps-in-time/. Alternative archived version of this page: https://web.archive.org/web/20260223125550/https://livingwithmachines.ac.uk/macromap-interactive-maps-in-time/.

34 For further discussion, see Giorgia Tolfo, 'Creating a National Picture: Digitising Ordnance Survey Maps', *Living with Machines* (blog), 16 January 2023, accessed 27 February 2026, https://livingwithmachines.ac.uk/creating-a-national-picture-digitising-ordnance-survey-maps; and Giorgia Tolfo, 'Georeferencing Ordnance Survey Maps', *Living with Machines* (blog), 16 January 2023, accessed 27 February 2026, https://livingwithmachines.ac.uk/georeferencing-ordnance-survey-maps/. Alternative archived versions of these pages: https://web.archive.org/web/20260227191315/https://livingwithmachines.ac.uk/creating-a-national-picture-digitising-ordnance-survey-maps/ and https://web.archive.org/web/20260227191529/https://livingwithmachines.ac.uk/georeferencing-ordnance-survey-maps/.

35 An interactive version of Figure 2.2 is included in the open-access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

36 For a more detailed account of these steps, see *MapReader Documentation*, accessed 18 February 2026, https://mapreader.readthedocs.io/en/latest/. Alternative archived version of this page: https://web.archive.org/web/20260219091958/https://mapreader.readthedocs.io/en/latest/index.html.

37 Chris Fleet, 'Creating, Managing, and Maximising the Potential of Large Online Georeferenced Map Layers', *e-Perimetron* 14, no. 3 (2019), https://www.e-perimetron.org/Vol_14_3/Fleet.pdf; Chris Fleet, 'Understanding User Needs: A Case Study from the National Library of Scotland', *Digital Preservation Coalition Technology Watch Guidance Note* (2022), https://www.dpconline.org/docs/technology-watch-reports/2774-understanding-user-needs-national-library-of-scotland/file.

38 *IIIF Maps Community*, accessed 19 February 2026, https://iiif.io/community/groups/maps/. The IIIF is a community-driven set of international open standards that enables cultural institutions to make their digital image and audio-visual resources more accessible and usable online. It uses standardised application programming interfaces (APIs) and data formats for creating interoperable viewing experiences. By providing a common digital language, IIIF allows users to seamlessly access, view and work with digital objects from different collections worldwide. The IIIF Maps Community group is working

specifically on improving features for presenting, annotating and reusing digitised map collections as IIIF resources. Contributions like the Georeference extension are enabling a range of new tools. See, for example, Martijn Meijers and Jules Schoonman, 'Mapping the Edge: A Novel Approach to Georeferencing Historical Map Series', *e-Perimetron* 20, no. 1 (2025): 1–51, http://www.e-perimetron.org/Vol_20_1/Meijers_et_al.pdf.

39 Details on saving predicted labels are available in the *MapReader Documentation*, accessed 18 February 2026, https://mapreader.readthedocs.io/en/latest/using-mapreader/step-by-step-guide/4-classify/infer.html#add-predictions-to-metadata-and-save. Alternative archived version of this page: https://web.archive.org/web/20260219092627/https://mapreader.readthedocs.io/en/latest/using-mapreader/step-by-step-guide/4-classify/infer.html#add-predictions-to-metadata-and-save.

40 McDonough et al., 'Reading Maps at a Distance', 300.

41 Evangeline Corcoran, Kasra Hosseini, Laura Siles, Smita Kurup and Sebastian Ahnert, 'Automated Dynamic Phenotyping of Whole Oilseed Tape (Brassica Napus) Plants from Images Collected under Controlled Conditions', *Frontiers in Plant Science* 16 (2025): 1443882, https://doi.org/10.3389/fpls.2025.1443882.

42 Daniel Wilson, 'Finding Words in Maps', *Living with Machines* (blog), 30 July 2019, accessed 18 February 2026, https://livingwithmachines.ac.uk/finding-words-in-maps/; Olivia Vane, 'Finding Words in Maps Part 2: Seeing the Results', *Living with Machines* (blog), 22 August 2019, accessed 18 February 2026, https://livingwithmachines.ac.uk/finding-words-in-maps-part-2-seeing-the-results/. Alternative archived versions of these pages: https://web.archive.org/web/20260223130245/https://livingwithmachines.ac.uk/finding-words-in-maps/ and https://web.archive.org/web/20260223125954/https://livingwithmachines.ac.uk/finding-words-in-maps-part-2-seeing-the-results/. Our code from these experiments can be found at https://github.com/jamespjh/strabo-text-recognition-deep-learning and is forked from Spatial-Computing/Strabo-Text-Recognition-Deep-Learning, C++, 14 February 2018; USC Spatial Computing Lab, released 10 June 2023, https://github.com/spatial-computing/strabo-text-recognition-deep-learning. Strabo is based on underlying research from Yao-Yi Chiang and Craig A. Knoblock, 'Recognizing Text in Raster Maps', *Geoinformatica* 19, no. 1 (2015): 1–27, https://doi.org/10.1007/s10707-014-0203-9. Alternative archived version of the LwM code: https://web.archive.org/web/20260223130456/https://github.com/jamespjh/strabo-text-recognition-deep-learning.

43 'Machines Reading Maps: Finding and Understanding Text on Maps' (AH/V009400/1), https://gtr.ukri.org/projects?ref=AH%2FV009400%2F1, was jointly funded by the National Endowment of the Humanities (US) and the Arts and Humanities Research Council (UK).

44 MRM data and metadata for the Rumsey collection is openly available via the Stanford Digital Repository: Machines Reading Maps, accessed 18 February 2026, https://purl.stanford.edu/rc349kh8402. Individual deposits include: Yao-Yi Chiang, Katherine McDonough, David Rumsey, and Zekun

Li, 'David Rumsey Map Collection Text on Maps Data [Version 1]', (Stanford: SDR, 2023), https://doi.org/10.25740/vn901vj0926; Yao-Yi Chiang, Katherine McDonough, David Rumsey, Zekun Li, 'David Rumsey Map Collection Text on Maps Data [Version 2]' (SDR, 2023), https://doi.org/10.25740/wc461hp2261. Metadata for this subset of the collection is also available in the same SDR collection: 'David Rumsey Map Collection: Georeferenced Maps Metadata' (Stanford Digital Repository, 2023), https://doi.org/10.25740/ss311gz1992. The underlying research for mapKurator has been published as Yijun Lin and Yao-Yi Chiang, 'Hyper-Local Deformable Transformers for Text Spotting on Historical Maps', in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2024): 5387–97, https://doi.org/10.1145/3637528.3671589. On annotating maptext, see Valeria Vitale, 'Map Text Annotation Guidelines', accessed 18 February 2026, https://github.com/machines-reading-maps/Tutorials-Newsletters/wiki/Map-Text-Annotation-Guidelines. Alternative archived version of this page: https://web.archive.org/web/20260223131005/https://github.com/machines-reading-maps/Tutorials-Newsletters/wiki/Map-Text-Annotation-Guidelines.

45 David Rumsey Historical Map Collection, accessed 18 February 2026, https://www.davidrumsey.com/. Alternative archived version of website from 14 January 2024, showing 'Text on Maps' search as originally designed on home page: https://web.archive.org/web/20240114125450/https://www.davidrumsey.com/.

46 Work from two projects using MapReader: Leon van Wissen, Manjusha Kuruppath, and Lodewijk Petram, 'Unlocking the Research Potential of Early Modern Dutch Maps', *European Journal of Geography* 16, no. 1 (2025): s12–17, https://doi.org/10.48088/ejg.si.spat.hum.l.wis.12.17; Robert K. Nelson, 'Mapping Environmental Inequalities: Using MapReader to Uncover Mid-Century Industrial Burdens', *Imago Mundi* 76, no. 2 (2024): 284–90, https://doi.org/10.1080/03085694.2024.2453333.

47 McDonough et al., 'Reading Maps at a Distance'.

48 Data/Culture. Building sustainable communities around Arts and Humanities datasets and software, https://gtr.ukri.org/projects?ref=AH%2FY00745X%2F1.

49 Katherine McDonough, Bert Spaan, Kaspar Beelen, Jules Schoonman, Valeria Vitale, Daniel C. S. Wilson and Rosie Wood, 'Text on Maps: From Collection Discovery to Research Data', paper presented at Fantastic Futures 2025, London, UK, 4 December 2025, https://doi.org/10.23636/54k0-ny43. The presentation is available online: Katherine McDonough and Bert Spaan, 'Text on Maps: From Collection Discovery to Research Data', accessed 23 February 2026, https://pie.yt/?v=https://youtu.be/_gh1mrrtmJQ?si=QF3AXz1DNo0D9qiT&t=45&pieshare=1.

50 Phillip C. Muehrcke, *Map Use: Reading Analysis and Interpretation*, 2nd ed. (JP Publications, 1986), vii.

51 See previously cited works by Edney and Harley, but also the History of Cartography project at the University of Wisconsin-Madison: J.B. Harley and

David Woodward, eds, *The History of Cartography*, 6 vols (University of Chicago Press, 1987–2019).

52 Denis Wood, with John Fels and John Krygier, *Rethinking the Power of Maps* (The Guilford Press, 2010), 39.

53 See McDonough et al., 'Reading Maps at a Distance', 296.

54 For further discussion, see McDonough, 'Maps as Data'.

55 Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (University of Illinois Press, 2011), 10, https://doi.org/10.5406/illinois/9780252036415.001.0001.

56 Ramsay, *Reading Machines*, 32.

57 Mark Algee-Hewitt, 'Computing Criticism: Humanities Concepts and Digital Methods', in *Computational Humanities* 35, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/138dbec6-12b4-4728-9ca9-6540ad095530.

58 Algee-Hewitt, 'Computing Criticism', 35.

59 Katherine Bode, 'The Archive', in *The Cambridge Companion to Literature in a Digital Age*, ed. Adam Hammond (Cambridge University Press, 2024): 106, https://doi.org/10.1017/9781009349567.007.

60 See for example, Robert Schwartz, Ian Gregory and Thomas Thevenin, 'Spatial History: Railways, Uneven Development, and Population Change in France and Great Britain, 1850–1914', *Journal of Interdisciplinary History* 42 (2011): 53–88, https://doi.org/10.1162/JINH_a_00205.

61 We have found that collective annotation is useful to establish basic guidelines, when the downstream analysis of MapReader output will involve more than one scholar. However, it was better for one person to annotate the final training data set to ensure that there were not conflicting versions of each label due to even minor inter-annotator disagreement.

62 Hannah Ringler adopts 'interpretive screen' when discussing Rockwell and Sinclair, *Hermeneutica: Computer-Assisted Interpretation in the Humanities* (MIT Press, 2016), https://doi.org/10.7551/mitpress/9522.001.0001. Ringler, 'Computation and Hermeneutics: Why We Still Need Interpretation to Be by (Computational) Humanists', in *Computational Humanities* 6, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/cdccb7af-e7cd-4225-932a-f12da4214b1a.

63 One might look for inspiration on next steps to the process that Jo Guldi outlines in her article 'Critical Search: A Procedure for Guided Reading in Large Scale Textual Corpora', *Journal of Cultural Analytics* 3, no. 1 (2018), https://doi.org/10.22148/16.030.

64 These image classification models in particular, we have found, do not always 'pay attention' to the areas of a patch that a human annotator might assume as most important. Using occlusion analysis and confidence scores side by side, we can study a model's decision.

65 Hosseini et al., 'MapReader', 6. F1 Scores capture a harmonic mean between precision and recall, expressed here as 2 x 'True Positive' patches divided by 2 x the sum of 'True Positive', 'False Positive' and 'False Negative' patches.

66 Kirsty Lilley, Katherine McDonough, David Alexander and Rosie Wood, 'Landscape Change and Conservation with MapReader in England's National Parks', poster presented at the CoSeC (Computational Science Centre for Research Communities) Annual Conference, Manchester, 3 December 2025. Using MapReader in HPC environments fulfils the prophecy of Quinn Dombrowski, Tassie Gniady, David Kloster, Megan Meredith-Lobay, Jeffrey Tharsen and Lee Zickel, in 'Voices from the Server Room: Humanists in High-Performance Computing', in *Computational Humanities*, 248, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/57bf0197-fd4a-44ce-a542-78883631cacb.

67 Katherine McDonough, Kaspar Beelen and Daniel C.S. Wilson, 'Neighbourhood Walks: A New Semantic Topology for Historical Map Text', paper presented at Computational Humanities Research 2025 (Lightning Talk Session), Luxembourg, 11 December 2025.

68 On the Peak District collaboration, see David Alexander, 'Mapping the Past against the Present', *ACID (Archaeology and Conservation in Derbyshire)* 23 (2026): 10–11, https://www.peakdistrict.gov.uk/__data/assets/pdf_file/0023/210578/ACID_Newsletter_2026.pdf.

69 *Allmaps*, accessed 18 February 2026, https://allmaps.org/; Rainer Simon, *Annotorious*, accessed 18 February 2026, https://github.com/annotorious; Eamonn Bell, Karina Rodriguez Echavarria and Jeyan Thiyagalingam, 'CCP-AHC Roadmap Open Draft', *Zenodo*, 11 September 2025, https://doi.org/10.5281/zenodo.17099176.

70 Leon van Wissen, 'Unlocking the Research Potential of Early Modern Dutch Maps'.

71 Among other works from the EPFL team, see Beatrice Vaienti, Isabelle di Lenardo and Frédéric Kaplan, 'Exploring Cartographic Genealogies through Deformation Analysis: Case Studies on Ancient Maps and Synthetic Data', *Cartography and Geographic Information Science* 52, no. 5 (2025): 557–77, https://doi.org/10.1080/15230406.2024.2424891; on Sanborn maps see Nelson, 'Mapping Environmental Inequalities'; on the MapText competitions see Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu and Nathalie Abadie, 'ICDAR 2024 Competition on Historical Map Text Detection, Recognition, and Linking', in *Document Analysis and Recognition – ICDAR 2024: 18th International Conference Proceedings*, Athens, Greece, 30 August to 4 September 2024 (Springer Verlag, 2024): 363–80, https://doi.org/10.1007/978-3-031-70552-6_22; Yijun Lin, Solenn Tual, Zekun Li, Leeje Jang, Yao-Yi Chiang, Jerod Weinman, Joseph Chazalon, Edwin Carlinet, Julien Perret, Nathalie Abadie, Bertrand Duménieu, Ta-Chien Chan, Hsiung-Ming Liao, Wen-Rong Su, Mengjie Zou, Tianhao Dai, Rémi Petitpierre, Beatrice Vaienti, Frédéric Kaplan, Isabella di Lenardo, Youngmin Baek, Michael Hentschel, Yu Nakagome, Ichimura Shuta, Jeongtae Lee and Chankyu Choi, 'ICDAR 2025 Competition on Historical Map Text Detection,

Recognition, and Linking', in *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025* (Springer Verlag, 2025), 568–85, https://doi.org/10.1007/978-3-032-04630-7_33.

72 Algee-Hewitt, 'Computing Criticism', 36.

73 Ahnert et al., *Collaborative Historical Research in the Age of Big Data*.

Chapter 3

Beyond the tracks: re-connecting people, places and stations in the history of late Victorian railways

Joshua Rhodes, Jon Lawrence, Kaspar Beelen, Katherine McDonough and Daniel C.S. Wilson[1]

Introduction

The drawn-out process of industrialisation first witnessed in Victorian Britain was ineluctably accompanied by urbanisation: either thanks to the gravitational pull of existing cities or in the creation of entirely new industrial ones. In the case of the former, cities such as Manchester, Newcastle, Birmingham and Glasgow offered sites in which heavy machinery could efficiently operate in proximity to its fuel sources (coal), as well as a disciplined and skilled labour force.[2] In the case of the latter, new manufacturing centres such as Middlesbrough, Barrow or Crewe announced themselves in different ways as exemplars of a new urban landscape. Transport, and the railway network, in particular, was at the heart of these transformations. Railways transformed the economics of agriculture by opening up new urban markets, and the spatial basis of industry itself by making fuel, raw material and finished goods mobile in new ways.[3] In doing so, they left their imprint on the shape of most towns and cities and influenced economic and social life across the whole country.[4] Railways scythed their way through built-up areas creating new geometric forms, separating neighbourhoods, clearing slums and creating new ones. They brought noise, vibration and smoke (once the preserve of the factory) into residential districts. But railways were also an immensely popular mode of passenger transport:

by the end of the century over a billion journeys a year were taking people to the places they needed or wanted to visit.[5] These incommensurable effects make the impact of a system like rail impossible to evaluate as a whole. The costs and benefits of a geographically distributed system do not always accrue equally – a problem that recurs with every infrastructural 'megaproject', and which has been dramatised in our day by disputes over the UK's planned high-speed railway, HS2. Local 'amenities' and 'disamenities' are rarely balanced against regional or national ones, and it has not hitherto been possible to aggregate these local effects in search of broader patterns. This chapter attempts to address this challenge as we bring together different forms of evidence for the first time to allow a new and more generalised understanding of the impact of railway infrastructure on people and places.

The physical extent of the railway, stretching as it does both within and between towns and cities, therefore makes the question of *scale* a key issue in any analysis.[6] The regional, inter-city and national character of the railway – typically understood as a network of connected points – has drawn the attention of economic and transport historians and historical geographers, whose interests, for example in measuring traffic flows or economic development, have encouraged a particular set of methods for working in aggregate.[7] These have predominated over the local or case-based approaches favoured by urban and social historians.[8] Histories of railways with people at their core have proven elusive, not only thanks to this methodological divide but also because important primary sources such as the census have, for technical reasons, been difficult to bring into conversation with other source materials relating either to transport, or to the spatial information found on maps. This chapter addresses some of these challenges by experimenting with interdisciplinary digital methods that allow new forms of convergence between source materials, combining detail about individuals at the micro-level, with analysis on a national scale.[9]

Although our focus in this chapter goes beyond urban areas, our method speaks to a long tradition of trying to understand the industrial city. Despite their almost irreducible complexity, cities have been seen as synonymous with human culture and society at large, and in a time of dramatic material change such as nineteenth-century Britain, the city often stood for industrial modernity itself. For this reason, investigators and analysts of different stripes have attempted to make sense of its new urban forms,

sensing that it offered a key to unlocking more than just the particular case study at hand.[10] When Friedrich Engels wrote his classic study of mid-nineteenth-century Manchester, the urgency came from his belief that this 'shock city' of the Industrial Revolution was a portent for the future of cities, and human society, everywhere.[11] Likewise, when attention turned to the slums of East London from the 1880s, it was because of a national outcry about the state of the nation, whose symptoms were found at their worst in its cities.[12] Campaigning investigators such as Arnold Toynbee, Beatrice Webb and Charles Booth (among others) joined Engels in producing work that today might be called geography or sociology: empirical but also theoretical, and nearly always using maps to put forward spatial reasoning and argumentation in visual form.

This desire to grasp the city in its shifting totality recurred in the self-styled science of 'civics' advocated in the early twentieth century by Patrick Geddes, for whom the shape of urban development could not be separated from its history. In considering the morphology of the world's then biggest example of the form, Geddes searched for an adequate metaphor: 'This octopus of London, polypus rather, is something curious exceedingly, a vast irregular growth without previous parallel in the world of life – perhaps likest to the spreadings of a great coral reef'. Using a new visual tool for combining population and transport on one map (see Figure 3.1) of this 'vast new unity', Geddes noted the 'new lines of union: the very word "lines" nowadays most readily suggesting the railways, which are the throbbing arteries, the roaring pulses of the intensely living whole'.[13] Contemporaries like Edward Thomas had no doubt that the railway took the curse of 'modernity' out into the countryside, but whether it was the cause or effect of suburban growth, it was undoubtedly the key to understanding it.[14] Geddes pleaded for contemporaries to take a unitary view of the whole, for which, as ever, he found existing frameworks lacking and coined what has become a keyword for the field:

> To focus these developments [. . .] of the geographic tradition of town and country in which we were brought up [. . .] we need some little extension of our vocabulary [. . .] Some name, then, for these city-regions, these town aggregates, is wanted. Constellations we cannot call them [. . .] what of 'Conurbations?' That perhaps may serve as the necessary word, as an expression of this new form of

> population-grouping, which is already, as it were subconsciously, developing new forms of social grouping . . .[15]

This suggestion of a causal relationship between spatial and social groupings became the focus of detailed attention, from the late 1950s, for a loose collection of scholars, including H.J. Dyos, John Kellett, Richard Dennis and Colin Pooley among others, centred on the Urban History Group, who sought to explore the connections – as David Cannadine put it – 'from shapes on the ground to shapes in society'.[16] A series of influential studies by historians made use of the work, on the one hand, of geographers, and on the other, of sociologists, to explore how social categories such as class and occupation 'had a clear spatial expression, and were reflected in distinctive social areas'.[17] Despite voluminous studies exploring residential differentiation in paradigmatic cases from a variety of theoretical and empirical perspectives, the field appeared to get bogged down in methodological and definitional issues.[18] To some extent this resulted from the divergent disciplinary commitments in play, but, equally, this is

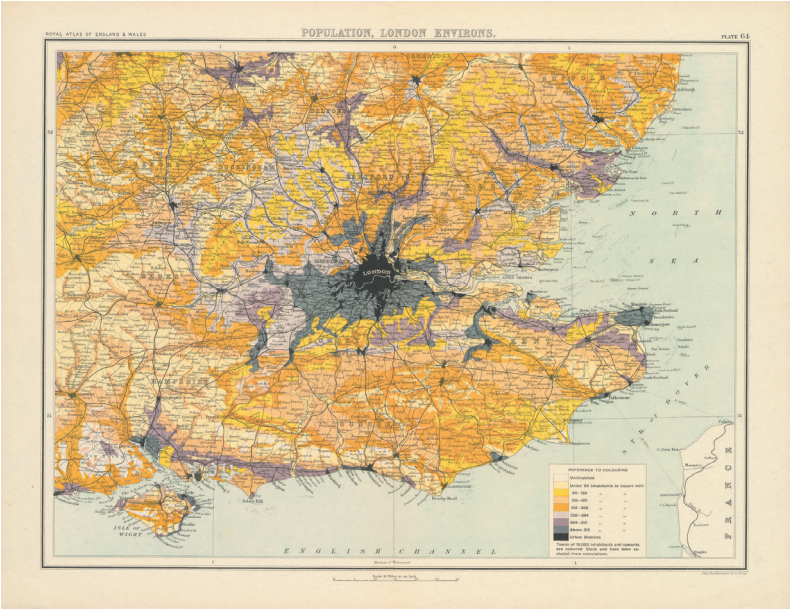


Figure 3.1: Map displaying both population density and railway lines, used by Geddes, (original from Bartholomew's *Royal Atlas of England*, 1899).

not so surprising, when we consider the highly heterogeneous nature of the data and evidence being used to make sense of supremely complex and over-determined processes of urban change.

John Kellett's masterly work on the Victorian city – to which we are indebted in this chapter – did more than most to clarify the impact of the railway on the five key cities that comprise his case studies; however, Kellett is cautious not to generalise unwisely from his examples.[19] The methods we outline in this chapter aspire, if not to generalise, then to propose a general framework for testing the effects of rail on residential differentiation nationwide, rather than just in large cities. Nonetheless, in so doing we remain mindful that the millions of individual actions that produced residential patterning were not always choices at all, and for most were choices constrained by economic and social factors; principally poverty, but also, for many, by dense networks of social obligation and reciprocity that underpinned their access to work, credit and, indeed, accommodation.[20] When, towards the end of *Ulysses*, Leopold Bloom is depicted imagining his perfect house beyond Dublin's city borders, James Joyce is reminding us that even for middle-class urbanites, unconstrained choice was the stuff of fantasy. It is, however, worth noting how Bloom, whose work required him to be a daily commuter, imagined his ideal location. His 'thatched bungalow shaped 2 storey dwelling house' would be set in extensive grounds and located 'not less than 1 statute mile from the periphery of the metropolis', and, perhaps most tellingly for our purposes, 'within a time limit of not more than 5 minutes from tram or train line'.[21]

Bloom is Joyce's mundane, if heroic, 'everyman', and so when we see him imagine the good life it remains anchored in the practicalities of daily life, including the fine calibrations of time and distance, amenity and disamenity.[22] In this regard, Bloom's outlook is strikingly different from that of the independently wealthy Margaret Schlegel in E.M. Forster's *Howard's End* (1910). Schlegel is portrayed viewing the rail network through the prism of social amenity rather than economic necessity. For her, the great railway termini of London represent 'our gates to the glorious and unknown. Through them we pass into adventure and sunshine, to them, alas! we return.'[23] Saved from the tiresome need to work for their living, the Schlegels, like many independently wealthy Edwardians, could view the railway network purely as a means of opening up the whole country to their pleasure. For them it was pure amenity; their use was occasional and they could live where they pleased, within reason, because they could

afford to be driven to the station when the need arose. The countervailing disamenity effects of rail have been acknowledged by urban and transport historians, but it has not previously been possible to explore these systematically alongside the rail network's amenity effects and to examine their possible interaction. Although financial and legal factors inclined railway companies to develop new urban lines (and their associated infrastructure) in poor, predominantly working-class districts, Kellett is clear that, in turn, these incursions could exacerbate the decline of blighted urban areas. At the local scale, dense rail infrastructure could reduce connectivity with surrounding areas, as well as bringing new environmental disamenities associated with steam power. Writing about London, Charles Booth described pockets of poverty being 'caught and held in successive railway loops', and Kellett shows that such 'barrier effects' operated in cities across Victorian Britain to cause, or exacerbate, the deterioration of urban neighbourhoods.[24] It is no accident that one of Gustave Doré's most iconic images of Victorian poverty represents densely packed working-class streets framed at either end by towering railway viaducts, and enveloped in the effluvia of a passing train. Such cultural touchstones offer useful texture in helping to situate our own account but also guide the hypotheses we seek to explore using more quantitative evidence, which necessarily operates at a smaller scale and with less qualitative detail. This balance between the close and distant reading of sources remains an open methodological question in our exploration of the human consequences of rail infrastructure as a feature of the changing landscape.[25]

As well as precipitating slum clearances, the incursion of rail into the city offered passengers a new perspective looking down on the cramped quarters of the urban poor.[26] This view became a familiar trope of Victorian literature that vividly encapsulated the class dynamics of slums and suburbs.[27] Clearly, there were many other sources of environmental disamenity in Victorian cities, especially for the poor. The deleterious effects of living close by rail barely feature in late Victorian poverty studies; industrial pollution and poor sanitation generally loom much larger. Railways also brought direct employment opportunities, especially for those in heavy industries and transport, which could make them a positive draw for some workers. Nonetheless, in the literature on urban history and town planning, rail is rarely conceptualised as an amenity, and its disamenity value remains highly subjective. In a seminal discussion, Dyos lists the types of amenity important in the establishment of Victorian suburbs: water, sewage, gas,

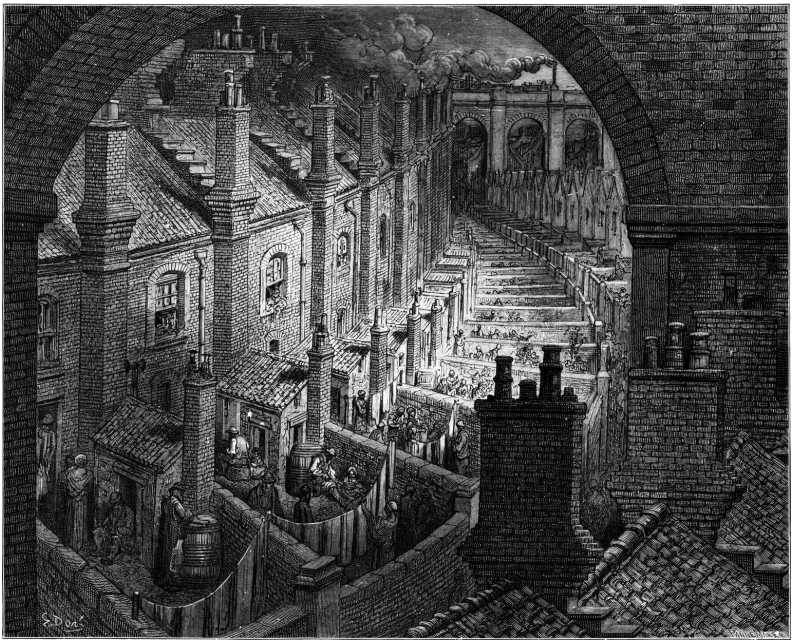


Figure 3.2: 'Over-London by rail', Gustave Doré print, from *London: A Pilgrimage* (Blanchard Jerrold, London, 1872). Science Museum Group Collection. © The Board of Trustees of the Science Museum. CC BY-NC-SA 4.0.

retail, recreation, churches and schools, but not railway stations.[28] With a few exceptions (such as in Glasgow or Nottingham which had interventionist municipalities) rail was conceived as something to which towns and cities were subjected by market forces. Conversely, an established legal framework existed for adjudicating harms caused by economic activities based on the concept of 'nuisance'. As Martin Daunton puts it:

> stopping a nuisance to one neighbour – say by an injunction against smells or noise from a workshop – could harm others by reducing employment. It became a matter of the balance of interests, and not absolute rights and wrongs. [. . .] A nuisance might be justified on the grounds that the activity had a general utility to society.[29]

But such arguments were rarely rehearsed in relation to railways, because the statutory basis on which companies had been granted access to land did not allow similar complaints to be made. Where such complaints could be

raised, this depended on an understanding of relevant legal provisions. For instance, letters to newspapers by well-informed middle-class correspondents cited the infringements of rules made decades earlier in relation to the considerate operation of steam engines; by contrast, the voices of those, perhaps more directly at the receiving end of rail's disamenity effects, living in the working-class districts, are rarely to be found in the archive. For this reason, among others, our notion of 'disamenity' is both subjective and qualitative. We infer that those in close proximity to railways could feel the sensory effects of the steam engine, including breathing its polluted air.

These effects impinged most heavily on the urban working class because it was only the better-off who could make choices about how to balance amenities and disamenities, the impacts of which were not evenly distributed. Many Victorian writers held that sensibility itself varied by social class. Explaining the feelings of a middle-class wife who refuses to move with her husband to poor lodgings, Gissing writes that she 'must have submitted to an extraordinary change before it would have been possible for her to live at ease in the circumstances which satisfy a decent working-class woman'.[30] We do not claim similar insight into how people *felt* about the environmental disamenities associated with dense rail infrastructure, but our evidence does suggest that the push and pull effects of rail worked differently for different social groups.

It was in Britain's great cities that the delicate balance between the amenity and disamenity effects of rail was hardest to strike, but our method goes beyond case studies that focus exclusively on the morphology of single cities and permits a larger scale of analysis.[31] By mapping the total physical imprint of railway infrastructure in conjunction with street-level demography, we uncover new types of spatial patterning in different residential contexts, thereby allowing new insights into who lived where and why. Realising this ambition has only been possible by converging datasets in wholly new ways.

Digital convergence: the view from the street

In this chapter, we combine three big, new and open datasets in a 'convergence experiment'. Each offers a different view of *fin de siècle* British society. The first dataset is digitised 1901 British census data that we have geo-coded at property and street level (*AddressGB*). The second is a dataset derived from historical Ordnance Survey (OS) maps of Britain that captures the location of

railway infrastructure (*Railspace*) around the turn of the twentieth century (see Chapter 2 for a full discussion of the MapReader tool used to create this dataset). The third contains the locations and details of railway passenger stations over time (*StopsGB*).[32] We created and restructured these datasets with an eye towards convergence, and establishing explicit spatial information for each source type laid the foundation for connecting them. The street operates as the central unit to which we link both socio-demographic data from the census and information about the railway network (proximity to passenger stations and rail infrastructure) we derive from the new *StopsGB* and *Railspace* datasets. Linking multiple datasets on a national scale at such a refined level of aggregation is a milestone for British historical research.

Although each dataset is derived from sources available online in the public domain in different ways, they were not always shared in ways that facilitated historical research using computational methods. There is, therefore, great novelty and utility in making them public in new ways. The census and maps were the products of major governmental initiatives during the Victorian period, which have been notoriously difficult to analyse in their entirety due to their size, complexity and problems of access. Our contribution to transforming the immensely complex, and sensitive micro census data is in adjusting the unit of analysis up from individuals (a level of data that cannot be shared publicly) and down from larger jurisdictions (such as parishes or registration sub-districts) to properties and streets. OS maps have been provided in different places and formats. We worked with the National Library of Scotland to access recently geo-referenced OS sheets through tile servers, still a relatively unique way of making big series map collections available online. As discussed in Chapter 2, processing these map images and creating new derived 'maps as data' objects is a big step towards breaking down big, unwieldy digital collections into usable historical data. Our third dataset, *StopsGB*, draws on the long-standing work of local historians and enthusiasts, previously available only in formats not amenable to computational methods. While each new dataset and the methods for their creation are described in more detail elsewhere, here we discuss the specific issues around data creation, aggregation, evaluation and linking that inform our experimental design.

Population census

A key component of our convergence dataset is geo-located street-level 1901 census data for Great Britain. We use Integrated Census Microdata

(I-CeM), a digitised version of the individual-level returns from decennial censuses of England, Wales and Scotland taken between 1851 and 1911.[33] I-CeM data contains the individual returns of everyone enumerated in the census. Table 3.1 is a snippet of the dataset, showing the Heighes family, who lived on Harrow Road in Paddington, London in 1901. The structure of I-CeM mirrors how the census was administered. Information was collected on individuals, who were grouped within households (or institutions, if they lived in a workhouse, for example). There were then a series of successively larger administrative units grouping these further: enumeration districts, parishes, registration sub-districts (RSDs), registration districts and registration counties.

Each row of I-CeM stores information for a single person. Individuals are listed according to their position in a household, reflecting the patriarchal structure of society, with men, typically a husband or father, listed as the head of the household. Everyone else enumerated in the household was listed in relation to the household head, for example, as wives, sons, daughters, boarders or servants. I-CeM also captures the rich information collected by census officials, including people's names, addresses, ages, gender, occupations and place of birth. In total, I-CeM contains over 100 variables for each person, combining information collected at the time and derived variables calculated after the digitisation and standardisation process.[34]

Despite the rich address information recorded in I-CeM, previous work has predominantly used aggregate administrative units (parishes, registration sub-districts, and counties) to map and spatially analyse the data.[35] Most recently, Bogart and colleagues have examined the macroeconomic impacts of improvements in transport infrastructure over the nineteenth century by linking the parish/RSD-aggregated census data to GIS transport datasets (rail, road and waterways).[36]

Yet the individual-level address data in I-CeM offers the possibility of spatial analysis at much higher resolutions than parishes and RSDs.[37] The challenge is linking the millions of textual descriptions of addresses in the census (e.g. 3 Harrow Road) to their 'real-world' locations. We have created an enhanced version of I-CeM, known as *AddressGB*, which links individuals and their street addresses in I-CeM to one of two existing GIS datasets through an automated Python pipeline.[38] The first dataset is OS Open Roads, a freely available and re-usable GIS of modern roads in Britain.[39] We use this modern roads dataset because there is no comparable

Table 3.1: Sample I-CeM data

RecID	First Name	Surname	Address	Parish	Sex	Age	Relationship	Occupation	Servants
14	Samuel	Heighes	3 Harrow Road	Paddington	M	33	Head	Coffee House Proprietor	4
15	Edith	Heighes	3 Harrow Road	Paddington	F	32	Wife		4
16	Samuel	Heighes	3 Harrow Road	Paddington	M	3	Son		4
17	Maud	Heighes	3 Harrow Road	Paddington	F	0.5	Daughter		4
18	Annie	Bailey	3 Harrow Road	Paddington	F	19	Servant	Barmaid In Coffee House	0
19	Agnes	Tarrant	3 Harrow Road	Paddington	F	26	Servant	Barmaid In Coffee House	0
20	Sarah	Hetherington	3 Harrow Road	Paddington	F	15	Servant	Nursemaid	0
21	George	Brooks	3 Harrow Road	Paddington	M	29	Boarder	Bath Attendant	0
22	Owen	Butler	3 Harrow Road	Paddington	M	35	Boarder	Bus Driver	0
23	Thomas	Youlden	3 Harrow Road	Paddington	M	26	Boarder	Carpenter	0
24	Frederick	Pain	3 Harrow Road	Paddington	M	24	Boarder	Carman	0
25	Walter	Eldridge	3 Harrow Road	Paddington	M	26	Boarder	Carpenter	0
26	Albert	Cowley	3 Harrow Road	Paddington	M	32	Servant	Porter	0

historical dataset of streets captured as line geometries for nineteenth- and early twentieth-century Britain. The second dataset is GB1900, a collection of geo-referenced, transcribed labels from the OS second edition County Series six-inch-to-one mile maps of Great Britain, published between 1888 and 1914.[40] The advantage of GB1900 is that it contains the historical names of streets (and sometimes individual properties), contemporaneous to our census data. The drawback is that GB1900 contains only point geometries, which refer to the upper-left corner of the text label on the map, and so it does not represent the actual shape or length of the road as OS Open Roads does.

We link individuals at street-level and not to the specific house or property in which they live. Individuals living at '3 Harrow Road' and '4 Harrow Road' will therefore all be linked to 'Harrow Road' in OS Open Roads and GB1900. Linking at street level has several advantages over linking to individual properties or houses. Geo-coding individual properties may be geographically more precise, but this precision may be historically meaningless. House numbers in the census may bear no relationship to modern house numbering, either because streets have been renumbered or because the numbers refer to the order in which properties were visited by the census enumerator. Streets also provide both much higher spatial resolution than parishes or RSDs and capture the 'character' of an area through its occupational structure and demography. Creating our new geo-coded datasets using property and street-level data retains a degree of anonymisation which enables us to publish our derived datasets open-source.[41]

Table 3.2 shows the number and proportion of people linked to each dataset. In total, we link 72.3 per cent of individuals to the street they lived on for the 1901 census. Overall, more people were linked to OS Open Roads than GB1900 but a substantial proportion of people (40 per cent) were linked to both OS Open Roads and GB1900. Linking rates were highest for urban areas, where the OS Open Roads coverage is closest to historical street topography, and weakest for remote rural areas, especially in Wales where matching problems have been compounded by shifts in the preferred spelling of Welsh place names. The overall linking accuracy was high, with less than 5 per cent of links identified as false positives through comparison with a stratified manual sample of addresses.[42] Including GB1900 data significantly improved linking rates outside urban areas and

Table 3.2: Number and percentage of individuals geo-coded in 1901 Great Britain Census

	Individuals	
Geo-coding Source	***N***	**%**
GB1900 Only	5,932,744	16
OS Open Roads Only	5,868,145	16
OS Open Roads and GB1900	14,826,561	40
Total Linked	26,627,450	72
Not Linked	10,303,062	28
All	36,930,512	100

also improved socio-economic representativeness nationwide, justifying their inclusion despite the limitations of their point data format.[43]

Reflecting the slight urban bias to the geo-coded dataset, younger people (under 40 years old) and women are slightly over-represented compared to the full population. Similarly, predominantly urban occupational sectors (most notably manufacturing) are over-represented and trades located outside urban centres (e.g. farming and estate work, and mining and quarrying) are slightly under-represented. We identified a particular over-representation of service, commercial and professional sectors, and an under-representation of general and factory labourers.[44] Our hypothesis is that this reflects the weaker linking of minor streets and courts compared with main roads and substantial residential streets (due partly to subsequent demolition of these minor streets, leading to their omission from OS Open Roads, and partly to the absence of labels for these streets on the historical maps transcribed for GB1900). Overall, we are confident that the greater precision made possible by linking most people to the specific street they lived on outweighs the disadvantages associated with no longer being able to make statements about the whole enumerated population (as can be done when resolving people to their parish of residence).

Railway infrastructure

Data from maps

Like the census, individual map sheets have long been used as sources by historians. But until recently it has not been possible to analyse

maps – especially series maps like the OS – at the collection level, for example, examining all sheets from one or more editions that have been digitised by the National Library of Scotland. In proposing to use the visual content of the large-scale national maps produced by the OS, we work against the grain of earlier urban history scholars. In their foundational study on British cities, for example, James Johnson and Colin Pooley depict maps as objective, sporadic sources.[45] On the one hand, maps could confirm the presence of specific features like a street or building unlike other, textual sources. However, because most city maps were surveyed and printed irregularly – both temporally, and spatially, in the sense that not all cities were covered equally – Johnson and Pooley set them aside as essential sources for a history of looking at all British cities. The latter concern strangely ignores OS topographic maps at slightly smaller scales. If OS Town Plans, printed at an exceptionally large scale (1:500), covered only urban centres and were rarely resurveyed or revised, the early editions of large-scale six-inch and twenty-five inch to one mile OS maps contain rich details about cities across Britain. With continuous coverage of urban, suburban and rural areas, these maps are key visual primary sources that contain evidence of industrial development. Recognising national OS mapping as a source for urban history means that we can shed the constraints faced by historians of the past, who visually compared limited physical maps across a small number of case studies. Our computational methods make it possible to transcend the comparative case study approach, by instead assessing all towns and cities simultaneously alongside one another, and in the context of their suburban and rural hinterlands.

But it is Johnson and Pooley's treatment of maps as containers of facts about the environment that any computational approach to maps must address head on. As detailed in Chapter 2, 'MapReader: Software and Principles for Computational Map Studies', such a view was typical until revisionist scholarship in the history of cartography which taught scholars to approach maps as we might any other historical source: as documents situated in their own contexts of production, intended and unintended uses, preservation status and power dynamics.[46] While these revisionist appeals for maps to be used as more than factual illustrations are now well known, applying such an approach at scale has never been done. Conscious of these barriers, in our work we treat the content of OS maps as a 'visual census' to be examined computationally alongside the population census: both have biases and histories that we must attend

to while analysing data derived from their digital versions.[47] Whether in an analogue or a computational context, maps cannot be treated with a positivist mindset. To make a critical, computational approach to maps as data possible, we developed MapReader, a software library that creates reproducible, labelled data based on queries of large collections of homogeneous maps.[48]

MapReader takes digitised map images and their metadata as input files, splits each image of a map sheet into regions called 'patches', and then uses computer vision classification models to produce output datasets that label patches based on researcher-defined concepts. A training dataset is created by manually categorising patches (in this case, sized as 100 m × 100 m areas) from a sub-sample of the map collection as 'rail' or 'no rail' depending on the presence of 'anything represented on the map related to the use of rail – including single and double tracks, stations, depots, and embankments'.[49] A computer vision model is then fine-tuned on these manual classifications of patches as 'rail' or 'no rail' and then used to classify patches across the whole map collection.

Taking an image classification approach at the patch level allows us to investigate visual phenomena present on thousands of maps relatively quickly compared to less efficient, more complex computer vision tasks (such as object detection or image segmentation). While the patch is an artificial spatial construct of squares running across and down the map, it is nonetheless sized by the researcher at the outset of an experiment in relation to a concept that has a visual signal on the map. And while patches are less precise than data captured as lines and points, using patch classification is highly effective at identifying areas containing information of interest. While annotating training data in MapReader, the researcher refines their guidelines for what counts as a 'rail' patch, and over time refines both their understanding of what is visually represented on maps and how it will relate to their research questions. Patches offer an intentionally coarse-grained approach whose benefits include reproducibility and speed: as such, they complement vector data and are a powerful tool when used to spot map features across large areas of the country with a level of precision that is suitable for many humanistic research purposes. MapReader's patch classification approach, therefore, lends itself well to making map content into machine-readable data for research.

In this chapter, we use a MapReader dataset of patches identifying rail infrastructure across all of Britain, hereafter the 'railspace' dataset. After

annotating roughly 62,000 patches manually,[50] we fine-tuned computer vision models accessible in the MapReader pipeline, selected the best performing one (resnest101e), and inferred the presence of rail infrastructure on 30.5 million patches with 96.74 per cent accuracy for every six-inch-to-one-mile OS maps printed around 1900 (the second edition sheets).[51] Out of this total, 487,360 patches or 1.3 per cent were classified as railspace.[52]

Our data is more comprehensive than other datasets representing historic railways because it contains tracks documented on maps that were not documented elsewhere.[53] It also captures the full spatial footprint of rail infrastructure, including tracks, embankments, associated buildings and any other visibly railway-related construction. This 'railspace' dataset offers a novel measure of how a key dimension of industrialisation impacted the physical, and therefore the social, economic and cultural landscapes of communities. Figure 3.3 demonstrates how this ambient, or environmental, effect of rail is made possible by MapReader's broader conception of

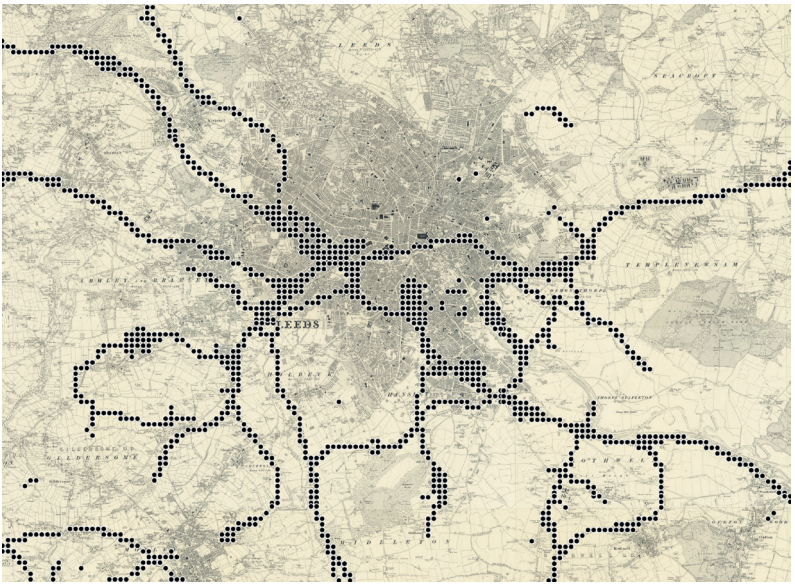


Figure 3.3: Detail of MapReader results for railspace patches (black) around Leeds. Credit: Katherine McDonough. Map image reproduced with the permission of the National Library of Scotland. CC BY.

railway infrastructure, that is, how its patches capture the overall footprint of rail infrastructure beyond its stations and routes (typically abstracted into points and lines in standard vector data representations).

Data from a database of passenger rail stations

Combining this enriched 'railspace' output with other sources on British rail infrastructure allows new questions to be posed. We use a newly created open dataset identifying over 12,000 railway passenger stations. Derived from Michael Quick's monumental community-sourced reference work *Railway Passenger Stations in Great Britain: A Chronology*, this information has been parsed and geo-coded to create a structured dataset suitable for spatial analysis.[54] The dataset includes information about station opening and closing dates, which has been used to exclude stations not operating in 1901.[55] In combination with MapReader's predictions about 'railspace', the *StopsGB* dataset allows exploration of the relationship between the railway network's overall footprint and its specifically passenger-facing aspects.[56]

The convergence experiments

Overview

We use the following method to combine our datasets, in each case classifying streets (and hence the people who lived on them) by their relationship to the rail datasets to investigate the interplay of rail and residential patterns as twin facets of industrialisation. First, we calculate the distance from each street in *AddressGB* to the nearest passenger station in *StopsGB*. This provides an indicator of proximity to railway stations as means of access to work or for general travel. Our analysis of proximity to stations uses this distance-to-station metric to group individuals, households and streets at various distance bands from stations, for example, *Under 250m*, *250–500m*, *500–750m*, *750m–1km* and so on. Next, we categorise streets by their proximity to substantial railway infrastructure (tracks, sidings, depots, etc.) as identified by MapReader. This provides an estimate of how strongly rail may have dominated the local environment. We define railspace proximity in terms of the percentage of MapReader patches classed as 'railspace' within 100 metres of each street.[57] Since the majority of people lived on streets where there was no railway infrastructure within 100 metres, we include 'No railspace' as its own category, while

streets near or within railspace are divided into bands from light (<20 per cent) to very dense (>80 per cent).

By combining our metrics for proximity to station and railspace we capture the different ways groups of individuals might use, interact with, and experience the rail network. For some, proximity to passenger stations offered amenity (the ability to traverse the rail network) but with varying degrees of disamenity depending on their proximity to heavy rail infrastructure and the associated nuisance of noise and pollution. For others, proximity to rail infrastructure, irrespective of distance to station, offered amenity through the employment opportunities that the railway brought. For this latter group, in contrast to rail network *users*, the amenity effect increased (at least in terms of employment opportunities) as railspace density increased.

People and stations

We began our analysis of the datasets with a number of hypotheses about the likely demographic patterns associated with proximity to passenger railway stations and dense rail infrastructure. We assume that *wealthier people* possessed the resources to be selective in where they lived, and therefore chose to locate themselves in areas perceived as 'desirable' within their social milieu. This *might* include access to transport networks, and avoidance of locations where transport infrastructure degraded the local environment. By contrast, poor and *disadvantaged people* are assumed not to have been able to be as selective in where they chose to live, and therefore, by default, lived in areas deemed undesirable by those with greater economic power. These locations *may* have been characterised by limited access to transport networks and/or disruptive impacts on the local environment as a consequence, among other things, of transport infrastructure and its associated trades. Finally, we assume that some people (as a consequence of employment) may derive particular amenity from easy access to transport networks and are therefore strongly motivated to live near transport nodes (including railway stations) despite some potential environmental disruption. Daily commuters to the central business district are the obvious example.

The census does not provide direct measures of either wealth or income, and so we have focused on demographic variables that provide surrogate markers, such as servant-keeping and occupation. According to Seebohm Rowntree, writing in 1901, employing live-in domestic servants represented

the key social (and cultural) marker of the 'middle class proper'.[58] Exploring such wealth-sensitive census variables in relation to residential proximity to passenger rail stations revealed clear differences between urban and rural parishes.[59] In urban areas, the percentage of households with live-in servants was highest close to stations (under 250 metres), falling sharply across distance bands between 250 m and 750 m, reaching its lowest point in the 1 to 1.25 km band, before rising again to a second, albeit lower, peak at 2 km from a station (beyond this the urban data become sparse and erratic, see Figure 3.4). By contrast, in rural parishes the proportion of households with live-in servants is highest at greater distances from a station (over 2 km), although here, too, values are relatively high close to a station, and there is a notable fall between 500 m and 1 km.

This allows us to draw important distinctions between the effects of rail infrastructure in urban and rural contexts, given that individuals living in urban areas were much more likely to live close to a station than their rural counterparts. In 1901, around 73 per cent of the national population lived in parishes defined as 'urban', and this preponderance of urban residents was even greater, at 86 per cent, among those living within 250 m of a railway station (see Appendix, Tables 3.4 and 3.5). By contrast, rural dwellers predominated among those living more than 3 km from a station, accounting for 74 per cent of households in this category (Table 3.5). Proximity to a station appears to have been much more important in urban and urban-fringe

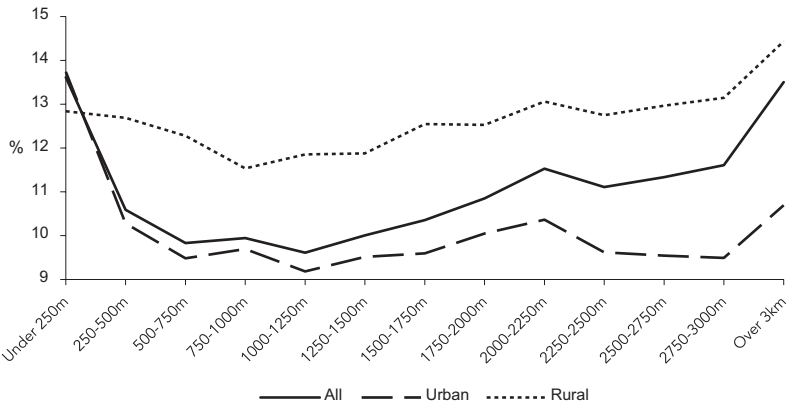


Figure 3.4: Percentage of households with live-in servants by distance to station and parish type, Great Britain (1901). Credit: Joshua Rhodes.

Table 3.3: Number and percentage of individuals by railspace density, 1901 Great Britain (within 100 m from street)

	Urban		Rural		All	
	N	%	*N*	%	*N*	%
Zero	13,374,693	73.92	4,718,719	26.08	18,093,412	68.07
< 20%	3,644,353	87.02	543,596	12.98	4,187,949	15.76
20–40%	1,885,075	85.35	323,552	14.65	2,208,627	8.31
40–60%	1,119,366	86.40	176,192	13.60	1,295,558	4.87
60–80%	524,566	84.47	96,448	15.53	621,014	2.34
> 80%	147,474	84.33	27,409	15.67	174,883	0.66

locations compared with rural areas, where distance to a station appears to exert a more muted influence on the residential patterns associated with high-status, servant-keeping households. This is consistent with the contrasting amenity uses associated with different factions of Rowntree's 'middle class proper', with the different worldviews, and practical needs, of daily commuters like Leopold Bloom compared with independently wealthy individuals like Forster's Schlegel family, for whom the railway network opened up an enlarged world of leisure and social connection.

The streetscape and railspace density

Alan Jackson's *Semi-Detached London* nicely captures how finely balanced calculations about the contrasting amenity and disamenity effects of rail shaped social segregation in the late Victorian suburb. In the typical suburb, he argues, '[M]ost of the houses would be within easy walking distance of the station, but the larger villas of the "carriage folk" were a little farther away, perhaps on higher ground. Nearer the station would be cottages of the serving classes and a row or two of shops with accommodation above for their owners.'[60] As Table 3.3 shows, in 1901 most people did not live within 100 m of rail infrastructure. Over 68 per cent of people had no railway tracks within 100 m of where they lived, and around 16 per cent lived near light rail infrastructure (defined as between 0 and 20 per cent of adjacent railspace patches). A further 16 per cent lived within railspace densities of over 20 per cent, but it was very rare for people to be surrounded by highly dense railspace (only 3 per cent of streets registered adjacent railspace densities at over 60 per cent). Individuals living in urban areas

constituted the overwhelming majority of people living near each band of railspace above zero – in all cases above 84 per cent.

Finally, Figures 3.5 and 3.6 show the percentage of servant-keeping households by both railspace *and* proximity to a station. Perhaps the most striking feature of these graphs is that they demonstrate that, just as Jackson suggests, it was relatively common to live close to a station, but away from the direct environmental influences of railway infrastructure ('zero' railspace). Interestingly, urban households display little difference in levels of servant-keeping between streets with no railspace and those with some (<20 per cent railspace). Indeed, in all the urban bands less than one

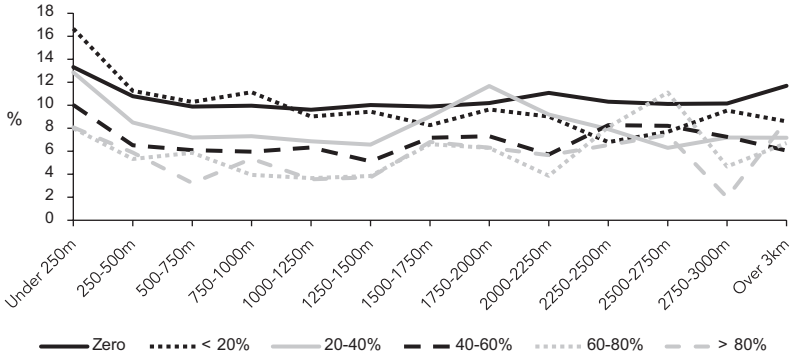


Figure 3.5: Percentage of urban households with live-in servants by distance to station and railspace density, Great Britain (1901). Credit: Joshua Rhodes.

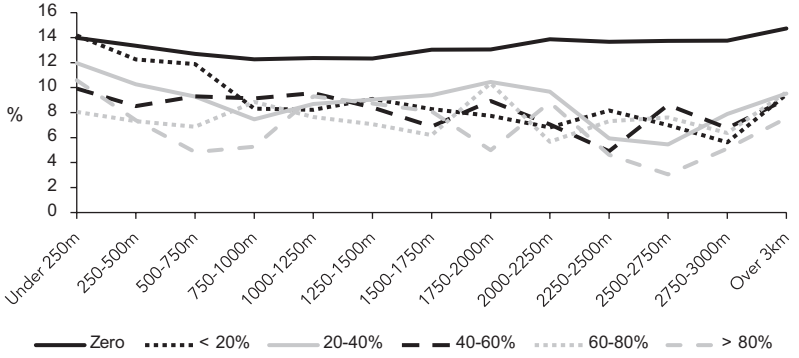


Figure 3.6: Percentage of rural households with live-in servants by distance to station and railspace density, Great Britain (1901). Credit: Joshua Rhodes.

kilometre from a station, it is streets close to some railspace, rather than those with none, that register the highest proportion of servant-keeping households (perhaps because these streets tended to be associated with better network access for the well-to-do daily rail user). It is also striking that, as hypothesised, urban streets surrounded by dense railspace (over 40 per cent) are associated with lower levels of servant-keeping, and that broadly speaking the denser the railspace, the fewer households with live-in domestic servants.

By contrast, the pattern is both different and less clear-cut among rural households. Here, except within 250 m of a station, households with no nearby railspace consistently register the highest values for servant-keeping, while levels of servant-keeping oscillate wildly for all other household types, especially beyond one kilometre from a station (Figure 3.6). Here the influence of proximity to a station on the location of wealthy, servant-keeping households appears much more muted. The main exception is households on streets within 250 m of a station, where, as in urban areas, servant-keeping households are slightly more common on streets with some nearby railspace (under 20 per cent), than on those with none. Equally, between 250 m and 750 m from a station there is a general tendency for most rural household types to register falling levels of servant-keeping households with increasing distance from the station. This tends to suggest that in rural, as in urban areas, there were some people for whom proximity to a station carried a significant advantage. It is possible that this reflects customer-facing establishments such as inns, hotels and retail premises.[61]

These patterns appear consistent with our hypothesis that, for well-to-do households, the amenity/disamenity calculus for living close to a station (and hence potentially also to rail infrastructure) would vary depending on how family members made use of the rail network (for daily work or occasional pleasure), and also how they travelled to the station (among Jackson's 'carriage folk' the need to live very near a station would have been less pressing, even for the daily commuter). To explore these differences further, and also to explore the non-network amenities that rail might bring to some groups of workers (through employment opportunities associated directly with rail infrastructure), we shift our focus to occupation. What type of worker was most likely to live close to a station but *not* close to dense rail? Did particular occupational groups tend to be

concentrated in areas of dense rail, or were such areas simply left to those with least choice over where to live?

Given the complexity of occupational classification systems, we develop our own occupational clusters using the finest-grained occupational codes from the digitised census.[62] We create occupational clusters for groups of workers who probably had distinctive relationships to either the amenity or disamenity aspects of rail. These consisted of four groups: workers whose employment was considered likely to be located in a central business district; workers directly employed on the railways, on railway building and maintenance or in making rolling stock; workers in heavy manufacturing and extractive industries likely to require supporting rail infrastructure, including the iron and steel and mining sectors; workers involved in the distribution and storage of goods.

Perhaps the most striking spatial pattern about the distribution of workers in these clusters was that the percentage of the workforce employed in heavy manufacturing increased as distance from the nearest passenger station increased (see Figure 3.7). If heavy industry was associated with rail transportation, this does not appear to have led to the creation of passenger rail stations, which seem unrelated to the location of workers in this sector. The other clear pattern to emerge was that, as predicted, the percentage of workers in occupations associated with the central

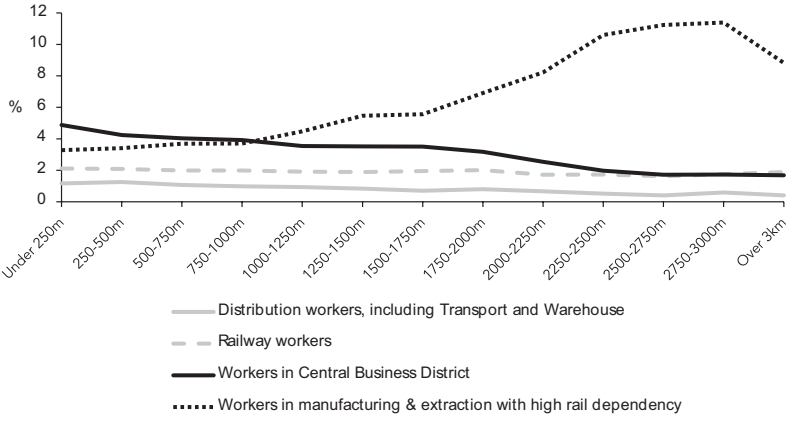


Figure 3.7: Percentage of all urban occupied workers in custom occupational groups by distance from nearest station, Great Britain (1901). Credit: Joshua Rhodes.

business district declined with increasing distance from the nearest station (Figure 3.7). It seems likely that these were the people most dependent on a daily commute for work: those for whom the amenity effects of rapid transit by rail were therefore greatest. Of course, there were many other ways to commute to work in late Victorian towns and cities, including omnibus, tram, bicycle and on foot.[63] But even so, the pattern for this group, which includes workers in banking, insurance, commerce and the principal professions, is clear. None of the other occupations appears to have a comparably strong association with proximity to railway stations.

But what happens when we add the density of rail infrastructure – our 'railspace' measure – to this story? Does this shed light on the perceived disamenity effects of living close by rail, and, if so, who felt them? How did wealthier individuals, those with most choice in the matter, negotiate the balance between amenity and disamenity, and can we identify groups for whom the economic benefits associated with proximity to rail infrastructure may have led them to be over-represented in areas of densest railspace? On a national scale, areas of densest rail infrastructure (over 60 per cent railspace density) tended to be associated with more workers occupied in heavy manufacturing, railway employment and goods distribution, and with fewer workers associated with occupations concentrated in central business districts (Figure 3.8). For many occupations, greater

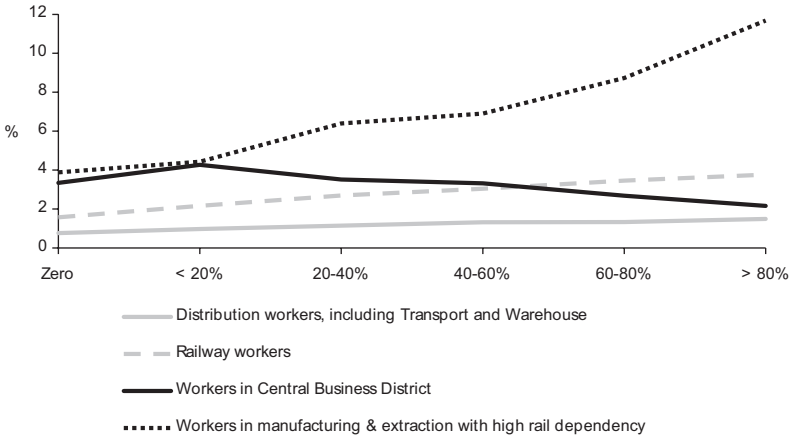


Figure 3.8: Percentage of workforce in custom occupational groups by railspace density, Great Britain (any distance from station), 1901. Credit: Joshua Rhodes.

railspace density was more strongly associated with the *pull* effects of employment in specific sectors dependent on rail, than with the push factors associated with railways as an environmental disamenity.

It does not follow that Gissing was right to claim working-class Victorians did not notice dirt and squalor, that their sensibilities were dulled. Economics alone is enough to explain why workers in some industrial sectors are more numerous in areas dominated by rail infrastructure. However, it is striking that only one of our occupational groups becomes markedly less numerous as rail density increased: workers associated with central business district occupations. We have already seen that these workers had a strong propensity to live close to passenger stations, and yet, their numbers fall at rail densities above 20 per cent. Since many workers in these occupations would also have been wealthier than average it may be that those with sufficient means to do so were making finely calibrated decisions about the balance of amenity and disamenity from rail.

The pattern is similar when one looks only at urban parishes and focuses just on those people living within 2 km of a railway passenger station (Figure 3.9). Again, the proportion of workers in central business district occupations rises between 0 and 20 per cent railspace density

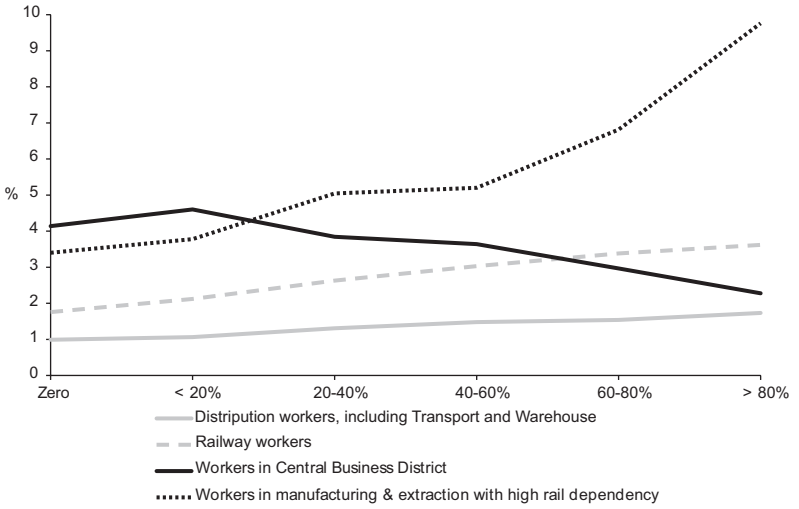


Figure 3.9: Percentage of workforce in custom occupational groups by railspace density (urban only; living less than 2 km from station), Great Britain, 1901. Credit: Joshua Rhodes.

and then declines. Indeed, if anything the fall is sharper, suggesting that dense rail infrastructure may have exerted a stronger push effect for central business district workers in urban contexts, while the patterns for all other groups are broadly similar to the national picture. As with the analysis of the distribution of servant-keeping households, this suggests that modest (as opposed to no) railspace may have offered many daily commuters the optimum trade-off between the amenity and disamenity effects of railways in late Victorian Britain. It also reminds us that these are patterns of association, not iron social rules. We still find central business district workers living in areas of very dense railspace, just as we also find servant-keeping households on such streets (see Figure 3.5), but their numbers are lower. Overall, this national-scale analysis of patterns between the distance to station, railspace density and the custom occupational groups highlights quite nuanced pictures for groups who had different desires, constraints and opportunities. It is the combination of the national coverage of the data and its geolocation at the street level that makes it possible to track these residential trends at such a fine level, thereby highlighting the different experiences of those working in city centres compared with those working in rail-related industries.

Bringing place back in

Having mapped these broad national patterns of association between street-level census data and rail infrastructure, we now turn to what was actually happening on the ground in specific places. We interrogate street-level patterns in four parishes where we already know something about the quotidian experience of life alongside Britain's rail infrastructure. These include two of London's well-known Victorian commuter suburbs – Camberwell and Upper Holloway – and two contrasting industrial districts – Salford (Lancashire) and Middlesbrough (Yorkshire). Harold Dyos' classic study of Camberwell, *Victorian Suburb* (1961), includes a detailed account of the impact of suburban rail development from the 1860s, as well as a case study of the Sultan Street district, where the 'barrier effects' caused by constructing new urban rail lines contributed to the 'creation of an urban slum'.[64] Upper Holloway is where George and Weedon Grossmiths' fictional City clerk Charles Pooter is said to live in his 'nice six-roomed residence'. At the opening of his diary, Mr Pooter describes how his family initially feared that noise from the railway at the bottom of their garden

would prove an inconvenience (and used this to secure a reduction on the rent). Pooter records that his landlord was right to argue that they wouldn't notice the trains after a while, noting: 'beyond the cracking of the garden wall at the bottom, we have suffered no inconvenience'.[65] It reminds us that people with social pretensions (Pooter has plenty) could learn to adapt to the inconveniences associated with living close by rail, although we should not forget that Pooter is a figure of (gentle) fun, for whom life is generally an unhappy compromise. As Figure 3.11 illustrates, the streets of Upper Holloway tended to be particularly close to passenger rail stations, although Pooter himself (like so many Londoners), relied on the cheaper omnibus service to commute back and forth to his job in the City.[66]

Middlesbrough and Salford, by contrast, were areas of heavy industry and docks, with vast areas devoted to the sidings and shunting yards required in an age when so much heavy freight needed to be moved by rail. In inner-city Salford, railways (and indeed waterways) had been punched through an already developed urban landscape, but in Middlesbrough rail infrastructure grew organically with the town during its great late Victorian iron boom.[67] Here, residential and industrial/transport sectors evolved in tandem and were therefore generally better segregated than across most of late Victorian Britain. Hence railspace density is only marginally higher for Middlesbrough than Holloway, whereas it is twice as high in Salford (Figure 3.10). We chose the 'Regent Road' parish of Salford because it included the cluster of terraced streets that Robert Roberts immortalised as 'the classic slum' in his semi-autobiographical account of the Edwardian working class.[68] Roberts famously described the thirty or so streets and alleys that delimited his childhood social world as being 'locked along the north and south by two railway systems a furlong apart' (see Figure 3.18).[69] Finally, the classic study of Middlesbrough in its industrial heyday is Lady Florence Bell's study *At the Works* (1907), which includes vivid accounts of the dangers trains and trucks posed to working-class children living on streets close to the town's great iron foundries; the segregation of railspace and residential space was far from absolute even when the two grew in tandem.[70]

Firstly, how does the rail density and proximity of passenger stations in our four case studies compare with other places? Figure 3.10 plots mean railspace density against mean distance to the nearest passenger railway station for 7,966 places across Britain, with our highlighted case studies.[71]

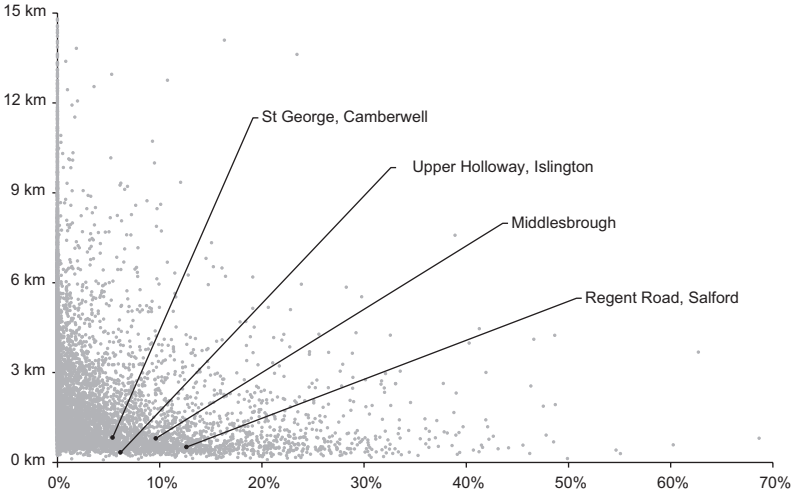


Figure 3.10: Mean railspace and distance to station by parish, Great Britain (1901). Credit: Joshua Rhodes.

Most places are clustered in the bottom left corner, reflecting the fact that, by 1901, the bulk of the British population lived *relatively* close to a station and on streets with little or no adjacent railspace. However, even at the parish level there was evidently considerable variation. As noted, railspace was significantly higher in Roberts' Salford than for the other three case study locations. But what can we learn by shifting our focus to visualising street-level patterns in the distribution of people and rail infrastructure? Are the broad patterns registered at national level amplified or obscured when visualised on the ground in our four case studies?

In the area of north London that includes Pooter's Upper Holloway, railspace was densest in the south and much sparser in the north (Figures 3.11 and 3.12). The type of rail infrastructure was different, too, with railspace areas in the south including extensive sidings and termini with warehouses and locomotive sheds (see the Midland Works and the Brickworks near Gospel Oak station in the bottom left, and the goods and coal depots to the bottom right). Both these areas have since been developed with the former now light industry and warehouses, and the latter the site for a Premier League football stadium. By contrast, the railspace to the north-east comprises an intersection of two lines. This consumes a smaller footprint and the nature of its use would have generated less noise, smoke

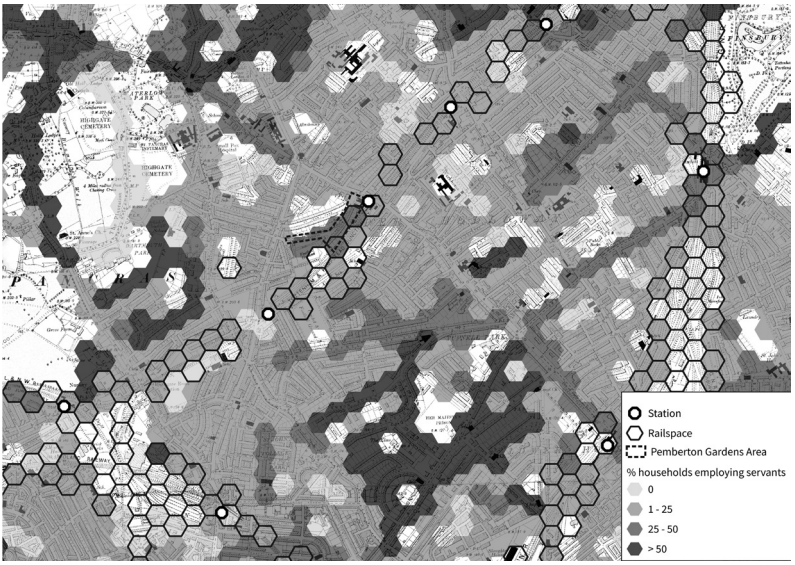


Figure 3.11: Map of spatial distribution of servant-keeping in relation to railspace and stations, Upper Holloway (London), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

and disruption than the dense rail infrastructure in the south. The streets in the middle of the intersecting railway lines in the top right still exist. It makes sense that servant-keeping was high in the north, across both non-railspace and railspace, but generally only high in the southern sections of the map in the middle area buffered from heavy railspace by streets dominated by households with lower levels of servant-keeping (Figure 3.11). Streets with the lowest levels of servant-keeping intersect with rail in many places in the south. The density and type of railspace is therefore a key factor in this residential patterning. The distribution of central business district workers follows a similar pattern to servants in the northern suburbs, but the highest band is widespread across the map, even close to the densest railspace in the south (Figure 3.12). Only in the bottom left (Gospel Oak) with its dense shunting yards was this less true. The differences in the distribution of servant-keeping households and central business district workers suggests that the servant measure is identifying subtle differences in social status among central business district workers. Though such workers are broadly present across most streets in north London, those with the means

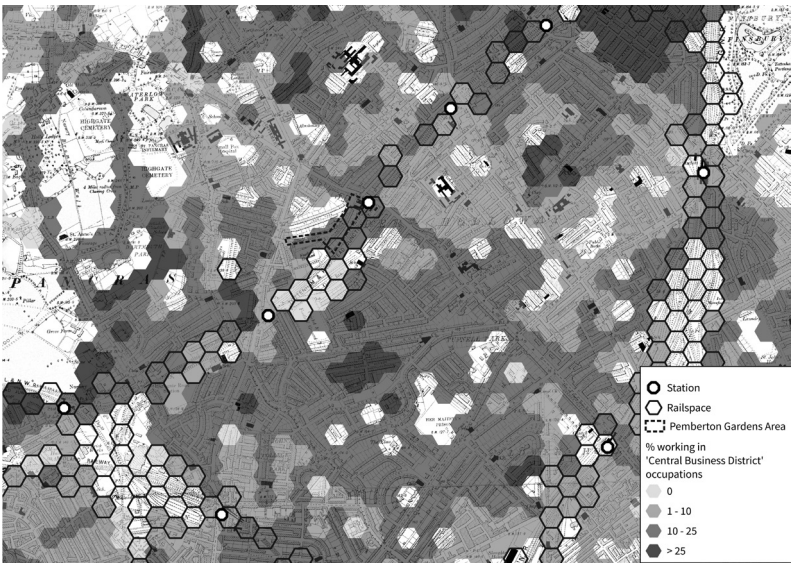


Figure 3.12: Map of spatial distribution of central business district workers in relation to railspace and stations, Upper Holloway (London), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

and space to employ live-in servants tended not to reside near the large rail sidings and goods sheds in the south.

The closest approximation to Pooter's fictional house 'The Laurels' may be located on Pemberton Gardens/Road (now all Pemberton Gardens), which is just to the west of Upper Holloway station. The gardens on the south side back on to the Tottenham and Hampstead rail line. Just under half the families on this street employed servants. Central business district workers made up just over a fifth of the 476 people living on the street. The houses got better towards Holloway Road (according to Booth), with the largest being up to ten to twelve rooms with a bathroom and 'one and often two servants kept'.[72]

One of the two most deprived streets in the area (according to Booth's reports) was located near the dense railspace running north–south from Finsbury Park to Holloway station. Queensland Road was right at the end of the sidings with a Great Northern Electric Light Station and a gas works. In the middle of Queensland Road, on the south side, was a Xylonite collar works, which produced plastic-coated shirt collars. Xylonite was a highly

flammable substance and early in the 1890s there was a serious fire at a factory manufacturing it at Homerton, East London. Although they had 'never had any fires here', such close proximity to potentially hazardous sites indicates the disamenities brought both by rail infrastructure and the industries it attracted.[73] The other heavily deprived street was Campbell Road, just west of Finsbury Park station. This was not a story of railway blight. Campbell Road was not hemmed in by rail, as Queensland Road was, and in fact was buffered from the rail lines and the Great Northern Goods and Coal Depot to the north of Clipton Terrace. Clipton Terrace itself was apparently showing 'signs of going down' by the 1890s, but a local police inspector could 'give no certain reason for the street [Campbell Road] having become so bad'.[74] According to Jerry White, 'Campbell Bunk', as it came to be known, became blighted because it was broken up into individual building plots, many of which languished undeveloped throughout the area's boom years, in turn making the whole street unattractive to the 'smaller servant-keeping class' for whom it was designed.[75]

In Camberwell, rail was generally less dense and mostly single or double rail lines except towards the north-east where there were sidings leading to a large number of goods sheds (just north of the image) and in the north-west, by the Thames, where there were four or more rail lines (Figure 3.13). Higher levels of servant-keeping were not located near these dense rail areas. The exceptions were main roads which, due to their retailing businesses and pubs, had higher servant-keeping across areas of both low and high railspace density. For example, the only street near the goods area in the north-east was the Old Kent Road, notoriously the most impecunious address on the Monopoly board. Residential side streets to the south had high levels of servant-keeping not found north of Camberwell New Road and Peckham Road. These streets were near stations and some railspace (but not near warehouses and sidings). As at Holloway, central business district workers were more ubiquitous than high servant-keeping households, but at Camberwell they were mostly concentrated in southern and western streets and clustered in areas between main rail lines but still close to stations (Figure 3.14).

The location of Dyos' slum includes Sultan Street and Hollington Street, both immediately to the west of the rail line running north–south in the centre. This railway, built in the 1860s, cut access to Camberwell Road, impoverishing these two streets, but by 1901 few streets in this area, except Camberwell Road itself, had significant numbers of servant-keeping households.[76]

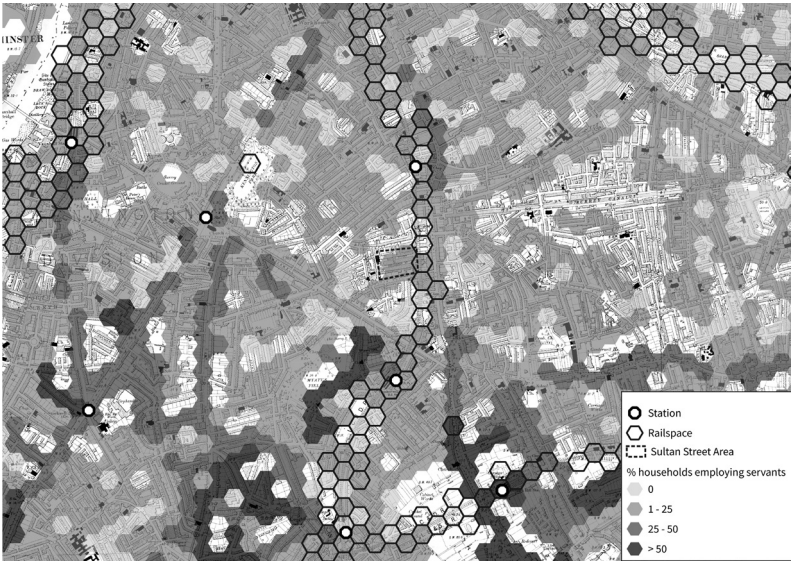


Figure 3.13: Map of spatial distribution of servant-keeping in relation to railspace and stations, Camberwell (London), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

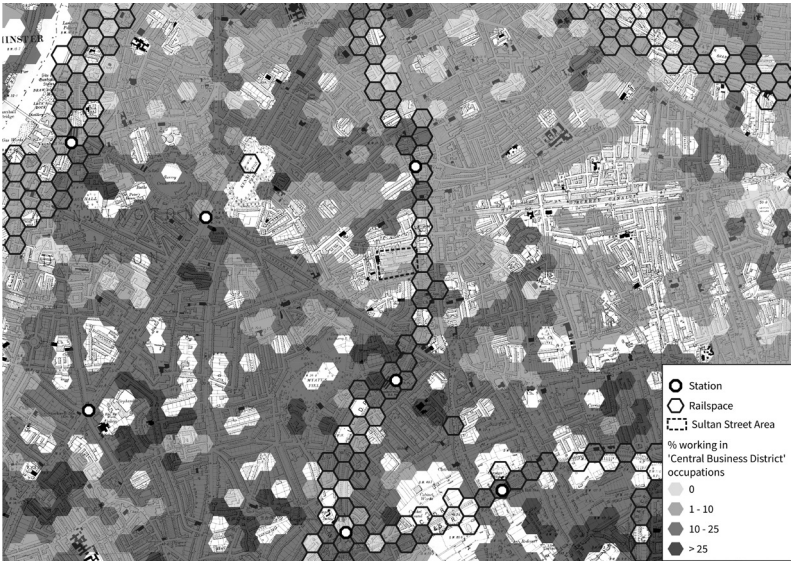


Figure 3.14: Map of spatial distribution of central business district workers in relation to railspace and stations, Camberwell (London), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

The unique situation of Sultan Street and Hollington Street may have made the coming of the rail line worse. One of Booth's assistants, George H. Duckworth, commented that 'these two streets are tucked away by themselves. There is a foot passage by the church out of the W[est] end of Sultan St. but none for carts'. The lack of cart access at that end was probably significant for limiting the disposal of waste (typically carried off in carts). Duckworth described Sultan Street as similar in 'appearance and character' to Hollington Street, which he noted had a 'fearful mess in [the] street, bread, meat, paper, vegetables, old tins [. . .] heavy rain but street full of children, bare head and bare feet [. . .] most windows broken'. The contemporary associations between deprivation and ethnicity are clear, with Duckworth noting that the inhabitants were 'Irish cockney general labourers not the carters of Beckett Street [the street below but with no direct link to Hollington Street]'. The backs were narrow and the one or two roomed houses lacked gardens, and there was 'much overcrowding'.[77] The railway cutting off access to Camberwell Road may have led to these streets' deprivation but by 1901 the main 'disamenities' were not the rail line to the east or distance from station (Wandsworth Road Station was less than 500 metres away). By the turn of the century, the significant environmental disamenity in this sub-district was the condition of the housing and overcrowding. The coming of rail may have tipped the balance, but at some point disamenities become mutually reinforcing and get baked in, as Jerry White shows for Islington's 'Campbell Bunk'.[78] The Sultan Street area experienced similar cycles of deprivation. These streets now form the Wyndham and Comber Estate.

In Middlesbrough the dense rail follows the river on both banks of the Tees but also cuts a swathe through the town centre heading towards Middlesbrough dock in the east. Streets with high rates of servant-keeping were scattered across the town, although they tended to be on main thoroughfares and were largely absent in the east around the North Ormesby area just south of the main dock (Figures 3.15 and 3.16). There were some high levels of servant-keeping in the northern part of Middlesbrough, surrounded by river and rail, but in general servant employers here were customer-facing concerns including hotel managers on Albert Street (north–south immediately above Middlesbrough station) and Innkeepers and Publicans on West Street (to the west of Market Place). Depot Road, the most northerly road with high servant-keeping extending into the heavy railspace area, had only two households – a railway platelayer and a foreman steam engine fitter. The foreman employed a general

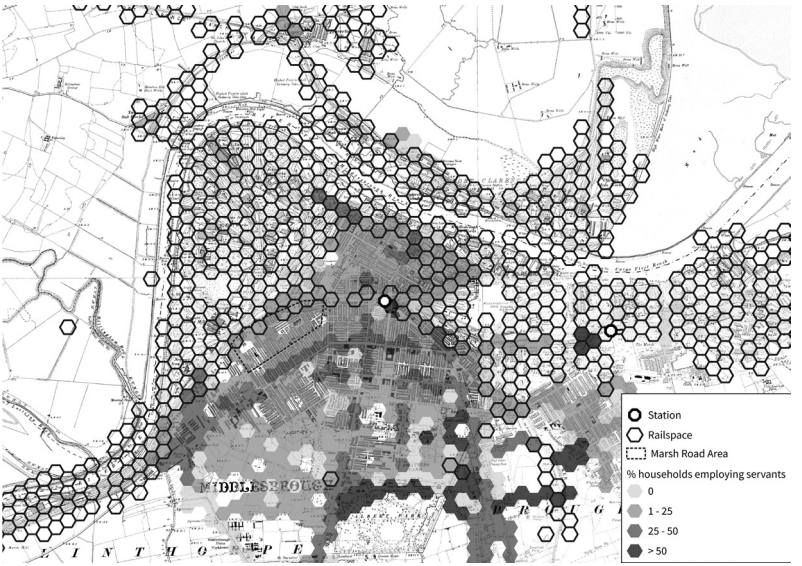


Figure 3.15: Map of spatial distribution of servant-keeping in relation to railspace and stations, Middlesbrough, 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

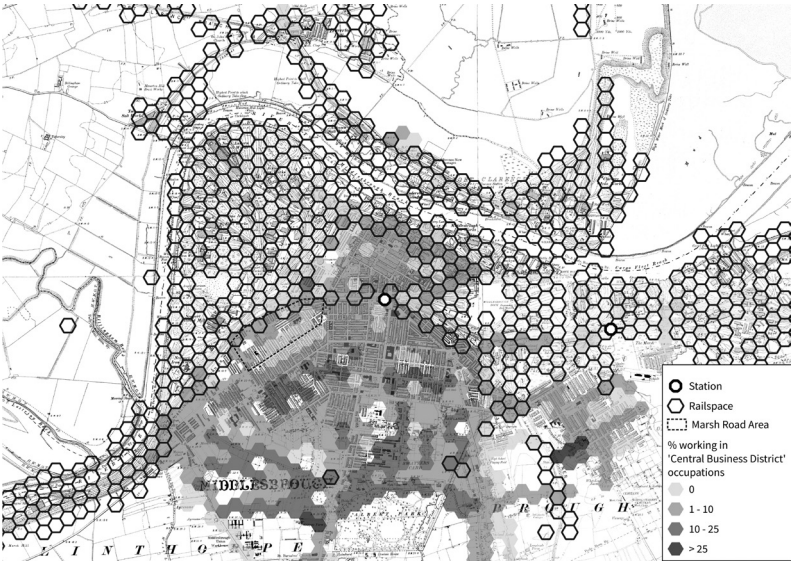


Figure 3.16: Map of spatial distribution of central business district workers in relation to railspace and stations, Middlesbrough, 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

domestic servant, accounting for the high (50 per cent) servant-keeping measure here. The distribution of central business district workers was much more clear-cut and shows them concentrated to the south of the town centre, especially far south near Albert Park, away from the dense railspace dominated by docks, sidings and warehouses. One apparent exception is Charlotte Street (around 14 per cent central business district workers) – located to the west above the railway line near Middlesbrough station – but there were only nine individuals living on the street, one of whom was a central business district worker. By living further from the dense rail infrastructure, central business district workers also distanced themselves from Middlesbrough station, but it is important to note that the compact geography of Middlesbrough and lack of rail stations reflect a town where most journeys to work would be on foot or bus. For workers with the means to choose where they lived, the push-pull factors of disamenity and amenity did not therefore always act in tandem. In the case of Middlesbrough, it was the push (away from dense railspace) that appears to have dictated the decisions of most well-to-do residents. There were, of course, other environmental disamenities influencing the desirability of certain locations. For example, the area between Newport Road and the mass of sidings to the east of Newport Rolling Mill regularly flooded in the 1890s and early 1900s. This included Marsh Road (railspace score of 0.49, indicating very high rail density nearby), where 'the land slopes towards the river and thus acts as a drain to the higher parts of the town'.[79]

At Salford, heavy railspace dominates around the docks in the south-west and south of the bend in the Irwell River (Figures 3.17 and 3.18). Roberts' 'classic slum' is located in the centre of the image, bounded by gas works to the west, sidings to the east and multiple mainline rail tracks to the north and south. There were few servant-keeping households except on Liverpool Street, a major thoroughfare with shops. Servant-keeping was generally low across the districts of Salford area between the two major clusters of rail infrastructure, with the exception of main roads such as Regent Road (which gives its name to the census registration sub-district covering this area). Higher rates of servant-keeping are generally only found in the suburban streets away from heavy rail infrastructure in the north-west and in the commercial heart of Manchester to the north-east. The distribution of central business district workers broadly follows servant-keeping with concentrations

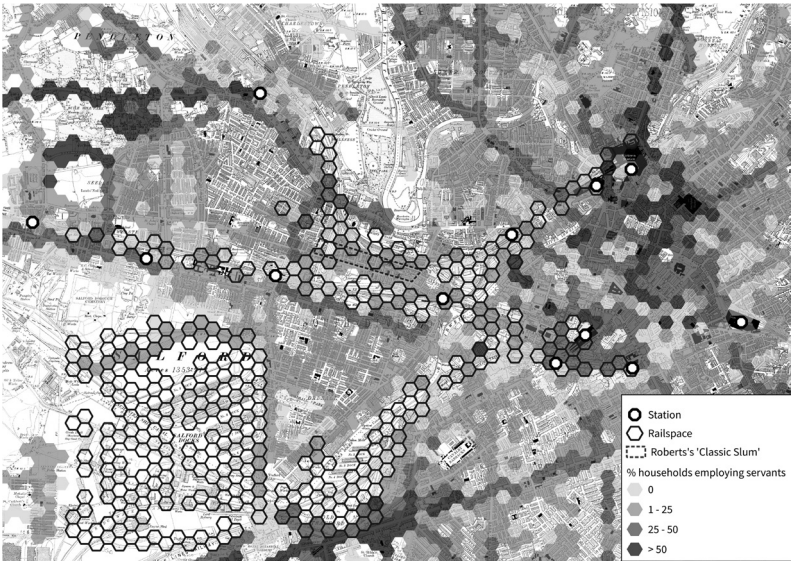


Figure 3.17: Map of spatial distribution of servant-keeping in relation to railspace and stations, Salford (Manchester), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

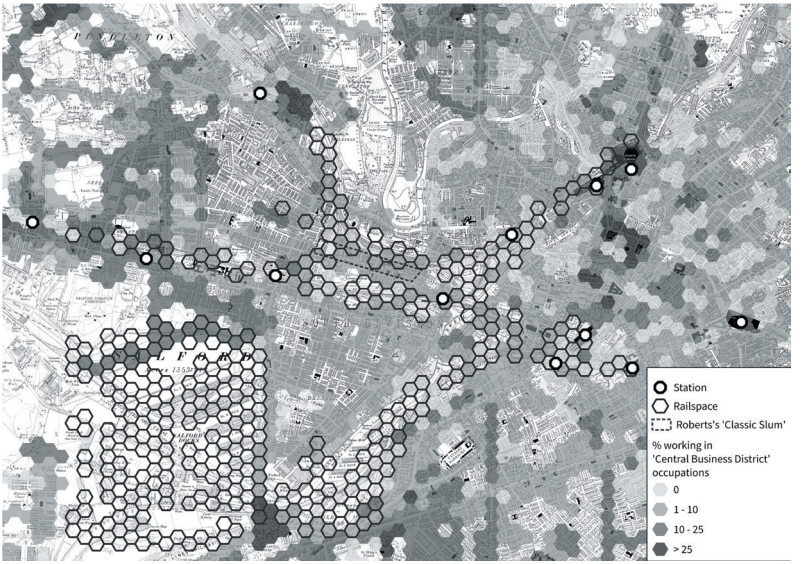


Figure 3.18: Map of spatial distribution of central business district workers in relation to railspace and stations, Salford (Manchester), 1901. Credit: Joshua Rhodes. Map image reproduced with the permission of the National Library of Scotland. CC BY.

in the western suburbs and central Manchester. They are absent from Roberts' classic slum area. Roberts was right when he recalled: 'West of us, well beyond the tramlines, lay the middle classes, bay-windowed and begardened. We knew them not.'[80] His 'classic slum' represents an extreme example of the disamenity effects that rail could bring. This was partly through 'barrier effects', where rail isolated the district from the surrounding cityscape, but this was doubtless reinforced when, as here, adjacent railspace was dominated by shunting yards and heavy industrial uses.

Conclusion

Structuring historically rich digital sources so that they can be converged in novel and flexible ways opens new perspectives on the social history of industrialisation and urbanisation. Our method allows us to shift between big-picture macro analysis, and the fine-grained, human-scale exploration of social context at street level. Consequently, we have been able to assess the impact of rail on the British population at the turn of the twentieth century in more detail than previously possible. In particular, we have illuminated the finely calibrated decisions of those able to make choices about where to live, including how proximity to rail infrastructure shaped these decisions. High-status households with servants chose to live close to stations, particularly in urban areas, and showed a clear preference for being situated within one kilometre of a station. These same households also avoided particularly dense rail infrastructure. But in urban areas it was hard to avoid rail infrastructure entirely, and, for some, light rail infrastructure had to be tolerated in exchange for the convenience of being near a station. These same patterns held true for certain types of occupational groups, such as those in trades associated with commuting to a central business district, but not for others, such as workers in manufacturing, whose relationship to rail infrastructure was very different. Workers in heavy manufacturing were in fact more likely to live further from stations than near to them but were disproportionately located in areas of high railspace, since although there were environmental disamenities, their livelihoods were tied to industries dependent on rail.

Our contrasting case studies of Camberwell, Holloway, Salford and Middlesbrough highlight the differential impact of railways depending on the type of rail infrastructure involved. At street-level, the spatial distribution of central business district workers and high-status servant-employing households mirrored national trends, with clear preferences for workers with the means to choose avoiding heavily industrialised areas with rail in preference for areas with no rail or little rail that were still near a station. The distinction between different types of railspace is important – extensive railway sidings and associated warehouses and factories represented a clear disamenity for those with means (albeit varying) to choose where they lived, but single or double rail tracks could be tolerated, even at close proximity. Holloway and Salford demonstrate most clearly the distinction between these two types of railspace and the character of the neighbours in the vicinity of each. Elsewhere, as in Camberwell, rail lines could exert a barrier effect that exacerbated perhaps already poorly connected streets thus leading to greater levels of deprivation than if rail had not been built nearby (notwithstanding the close proximity of stations). As Charles Booth's notebooks underline, neighbourhoods near heavy railspace also display other markers of deprivation (e.g. overcrowding) that we have not considered in detail in this chapter, but that were clearly important environmental disamenities. There is more that could be explored about the relationship between pre-existing deprivation, how it was exacerbated by the advent of different types of rail infrastructure, and how the two intersected in rendering certain streets and neighbourhoods deprived.

This chapter is very much a first word on these themes using new data. There are many avenues for further exploration. Possibilities include considering the persistence of deprivation today in the same areas highlighted in our case studies. This raises the opportunity to use *AddressGB* to look at historical antecedents of modern deprivation. It would also make it easier to explore how environmental disamenity interacted with ethnicity as well as class, since modern data collections register ethnicity in a way that historic census data does not (with historic data we only have information on place of birth). It would also be fruitful to identify new case studies either by using both the parish aggregated data to explore outliers, or by applying clustering techniques to identify streets across

Great Britain that shared similar characteristics and relationships with rail infrastructure (patterns that would otherwise not be identifiable manually). There is also great scope to use these methods to explore change over time since all three sources (*AddressGB*, *StopsGB* and OS maps) have a temporal dimension.

Future MapReader experiments, with new training data and models, might also capture the tram, as well as the rail network, as the two systems evolved in urban and suburban areas. This would make it possible to compare both systems at the national scale for the first time. Bloom's indifference to whether his rapid transit into central Dublin would be secured by train or tram probably *was* typical. It is widely recognised that the combined development of the municipal train and suburban rail systems played a decisive role in the social transformation of British cities, increasing levels of residential segregation by income/class and facilitating rapid urbanisation. This gave life to Geddes' neologism 'conurbation' as a way for contemporaries to describe their new regional mega-cities.[81]

We openly release our code and data to ensure the reproducibility of the findings in this chapter and to support the development of future work.[82] Throughout this research, we have made decisions that facilitate such openness. We hope that such open research practices for historical research will become standard, ensuring that historians can critically engage with the thought processes and interpretative decisions made throughout the processes of data creation, curation and analysis.

Appendix

Table 3.4: Number and percentage of individuals by urban/rural classification, Great Britain, 1901

	Urban		Rural		All
	N	%	*N*	%	*N*
England & Wales	24,480,229	75.34	8,012,379	24.66	32,492,608
Scotland	2,508,208	56.51	1,930,210	43.49	4,438,418
Great Britain	26,988,437	73.08	9,942,589	26.92	36,931,026

Table 3.5: Number and percentage of individuals at different distances from stations by urban class

	Urban			Rural			All
Distance from station	N	% (of distance band)	% (of all Urban)	N	% (of distance band)	% (of all Rural)	N
Under 250 m	3,703,973	86.33	17.80	586,552	13.67	9.93	4,290,525
250–500 m	4,914,843	87.12	23.62	726,607	12.88	12.30	5,641,450
500–750 m	4,161,769	87.39	20.00	600,587	12.61	10.16	4,762,356
750–1000 m	2,764,681	86.33	13.29	437,939	13.67	7.41	3,202,620
1000–1250 m	1,693,940	83.84	8.14	326,534	16.16	5.53	2,020,474
1250–1500 m	1,013,691	79.34	4.87	263,954	20.66	4.47	1,277,645
1500–1750 m	683,879	74.04	3.29	239,758	25.96	4.06	923,637
1750–2000 m	469,360	67.85	2.26	222,368	32.15	3.76	691,728
2000–2250 m	281,377	56.94	1.35	212,759	43.06	3.60	494,136
2250–2500 m	232,498	52.20	1.12	212,929	47.80	3.60	445,427
2500–2750 m	164,697	47.85	0.79	179,527	52.15	3.04	344,224
2750–3000 m	126,725	42.78	0.61	169,510	57.22	2.87	296,235
Over 3 km	593,989	25.55	2.85	1,730,544	74.45	29.28	2,324,533
All	20,805,422		100	5,909,568		100	26,714,990

Table 3.6: Custom I-CeM occupational coding

Rail Dependence	Occode
Central Business District workers	5, 6, 8, 9, 38–43, 48, 53, 56–62, 72, 78, 79, 97, 111–116, 118–124, 780, 782–784
Distribution workers, including transport and warehouse	34, 80, 114, 146, 162, 163, 166–170, 643, 730
Workers in manufacturing and extraction with high rail dependency	64, 142, 147, 196, 198, 200–203, 207–250, 252, 259, 260, 264, 334, 468
Railway workers	125–130, 132–134, 199, 265, 281, 360, 429, 432

Notes

1 Lawrence, McDonough, Beelen and Wilson conceived the original idea to 'converge' datasets to study place and railway. All authors collaboratively developed the research design and jointly conceived the data analysis. The bulk of the manuscript was written by Rhodes, Lawrence, McDonough and Wilson, with all authors contributing to revising and editing the final text. Rhodes produced the tables and figures.

2 Andreas Malm, *Fossil Capital: The Rise of Steam Power and the Roots of Global Warming* (Verso Books, 2016), especially chapter 7; E.A. Wrigley, *The Path to Sustained Growth: England's Transition from an Organic Economy to an Industrial Revolution* (Cambridge University Press, 2016).

3 David Turnock, *An Historical Geography of Railways in Great Britain and Ireland* (Ashgate, 1998).

4 A.K.B. Evans and J.V. Gough, eds, *The Impact of the Railway on Society in Britain: Essays in Honour of Jack Simmons* (Ashgate, 2003).

5 B.R. Mitchell, *British Historical Statistics* (Cambridge University Press, 1988), 547.

6 Daniel C.S. Wilson 'Working at Scale: What Do Computational Methods Mean for Research Using Cases, Models and Collections?', *Science Museum Group Journal* 18 (2023), accessed February 13, 2026, https://dx.doi.org/10.15180/221805.

7 See Dan Bogart, Xuesheng You, Eduard J. Alvarez-Palau, Max Satchell and Leigh Shaw-Taylor, 'Railways, Divergence, and Structural Change in 19th-Century England and Wales', *Journal of Urban Economics* 128 (2022): 1–23, accessed February 13 2026, https://doi.org/10.1016/j.jue.2021.103390; Stephan Heblich, Stephen J. Redding and Daniel M. Sturm, 'The Making of the Modern Metropolis: Evidence from London', *Quarterly Journal of Economics* 135, no. 4

(2020): 2059–133, accessed February 13, 2026, https://doi.org/10.1093/qje/qjaa014. For a spatial approach see, Robert Schwartz, Ian Gregory and Thomas Thévenin, 'Spatial History: Railways, Uneven Development, and Population Change in France and Great Britain, 1850–1914', *The Journal of Interdisciplinary History* 42, no 1 (2011): 53–88, accessed February 13, 2026, https://www.jstor.org/stable/41291135.

8 In nearly five decades, the journal *Urban History* has published only one article on the subject of Britain's railways; see Simon T. Abernethy, 'Opening up the Suburbs: Workmen's Trains in London 1860–1914', *Urban History* 42, no. 1 (2015): 70–88, accessed February 13, 2026, https://www.jstor.org/stable/26398343: an article which stands out by connecting transport and demography.

9 Likewise, we use literary examples to illustrate culturally resonant questions, but whose resolution 'at scale' can nonetheless only be attempted by moving to a more empirical register more typical of historical enquiry. These questions of shifting between both form and scale are addressed in Katherine Bode, 'The Equivalence of "Close" and "Distant" Reading: Or, Toward a New Object for Data-Rich Literary History', *Modern Language Quarterly* 78, no. 1 (2017): 77–106, accessed February 13, 2026, https://doi.org/10.1215/00267929-3699787.

10 Monika Krause, *Model Cases: On Canonical Research Objects and Sites* (University of Chicago Press, 2021), see chapter 1 on changing predilections among American scholars in the twentieth century for focusing on Chicago or Los Angeles as emblematic of the 'urban', as well as the consequences. For greater detail on the former, see Thomas F. Gieryn, 'City as Truth-Spot: Laboratories and Field-Sites in Urban Studies', *Social Studies of Science* 36, no. 1 (2006): 5–38, accessed February 13, 2026, https://doi.org/10.1177/0306312705054526.

11 Friedrich Engels, *The Condition of the Working Class in England in 1844* (Allen & Unwin, [1845] 1892), especially chapter 2.

12 E.P. Hennock, 'Concepts of Poverty in the British Social Surveys from Charles Booth to Arthur Bowley', in *The Social Survey in Historical Perspective, 1880–1940*, ed. Martin Bulmer, Kevin Bales and Kathryn Kish Sklar (Cambridge University Press, 1991), 189–216.

13 Patrick Geddes, *Cities in Evolution: An Introduction to the Town Planning Movement and to the Study of Civics* (Williams and Norgate, 1915), 26–7.

14 Edward Thomas, *In Pursuit of Spring* (Little Toller Books, [1914] 2016), 46–8, 50–1.

15 Geddes, *Cities*, 34.

16 David Cannadine, 'Residential Differentiation in Nineteenth-Century Towns', in *The Structure of Nineteenth Century Cities*, ed. James Henry Johnson and Colin G. Pooley (Croom Helm, 1982), 235–51.

17 Cannadine, 'Residential Differentiation', 237 (citing Colin Pooley).

18 Cannadine in 1982 lamented 'the problem of the historian when he [*sic*] wheels his trolley around the supermarket of ideas in search of theoretical packages or bodies of data from other disciplines which may help him order his own

inchoate mass of evidence: namely that in most other disciplines, from whose findings and theories he may wish to borrow, there is no agreement as to just what those theories and findings are', Cannadine, 'Residential Differentiation', 238.

19 John R. Kellett, *The Impact of Railways on Victorian Cities* (Routledge and Kegan Paul, 1969); the five cities are London, Birmingham, Manchester, Liverpool and Glasgow.

20 Gareth Stedman Jones, *Outcast London: A Study in the Relationship between Classes in Victorian Society* (Clarendon Press, 1971).

21 James Joyce, *Ulysses* (Penguin, [1922] 2000), 837–8.

22 According to Alan Jackson, London estate agents and transport executives agreed that the optimum catchment area around suburban stations was the 'magic half-mile': Alan Jackson, *Semi-Detached London: Suburban Development, Life and Transport, 1900–39* (Allen & Unwin, 1973), 220–21, 241–2.

23 E.M. Forster, *Howard's End* (Penguin, [1910] 1989), 27.

24 Kellett, *Impact*, 293–5; Richard Dennis, *English Industrial Cities of the Nineteenth Century: A Social Geography* (Cambridge University Press, 1984), 127–32; David A. Reeder, *Charles Booth's Descriptive Map of London Poverty 1889* (London Topographical Society, 1984), citing Charles Booth, *Life and Labour of the People in London*, 3rd ser. (Macmillan, 1902), 1–192.

25 See Bode, 'The Equivalence of "Close" and "Distant" Reading'.

26 On the view from the railway carriage in general, see Wolfgang Schivelbusch, *The Railway Journey: The Industrialization of Time and Space* (Berg, 1986); on slum clearances, see Kellett, *Impact*, chapters 10 and 11, and witness Friedrich Engels: 'Of late, the Liverpool railway has been carried through the middle of them, over a high viaduct, and has abolished many of the filthiest nooks; but what does that avail? Whoever passes over this viaduct and looks down, sees filth and wretchedness', Engels, *Condition of the Working Class*, 92.

27 H.J. Dyos and D.A. Reeder, 'Slums and Suburbs', in *The Victorian City*, ed. H.J. Dyos and M. Wolff (Routledge and Kegan Paul, 1973), 359–86. For example, Charles Booth, 'Condition and Occupations of the People of East London and Hackney, 1887', *Journal of the Royal Statistical Society* 51, no. 2 (1888): 282, accessed August 19, 2026, https://doi.org/10.2307/2979109. George Gissing, *The Nether World* (Penguin, [1889] 1973), 164.

28 H.J. Dyos, *Victorian Suburb: A Study of the Growth of Camberwell* (Leicester University Press, [1961] 1973), 138–68.

29 Martin Daunton, *State and Market in Victorian Britain: War, Welfare and Capitalism* (Boydell Press, 2008), 92.

30 George Gissing, *New Grub Street* (Penguin, [1891] 1985), 179.

31 Kasra Hosseini, Katherine McDonough, Daniel van Strien, Olivia Vane and Daniel C.S. Wilson, 'Maps of a Nation? The Digitized Ordnance Survey for New Historical Research', *Journal of Victorian Culture* 26, no. 2 (2021): 284–99, accessed February 13, 2026, https://doi.org/10.1093/jvcult/vcab009.

32 Mariona Coll Ardanuy, Kaspar Beelen, Jon Lawrence, Katherine McDonough, Frederico Nanni, Joshua Rhodes, Giorgia Tolfo, and Daniel C.S. Wilson, 'Station to Station: Linking and Enriching Historical British Railway Data', *Computational Humanities Research Workshop Proceedings* (2021): 249–65, accessed 13 February 2026, https://bl.iro.bl.uk/concern/conference_items/ff6f8bf8-03b8-4d00-824d-e786d54eae73 and the dataset itself: Mariona Coll Ardanuy Kaspar Beelen, Jon Lawrence, Katherine McDonough, Frederico Nanni, Joshua Rhodes, Giorgia Tolfo, and Daniel C.S. Wilson, 'StopsGB: Structured Timeline of Passenger Stations in Great Britain' (2021), accessed 13 February 2026, https://doi.org/10.23636/wvva-3d67.

33 K. Schurer and E. Higgs, *Integrated Census Microdata (I-CeM), 1851–1911* [data collection] (UK Data Service, 2020) SN: 7481, accessed 13 February 2026, http://doi.org/10.5255/UKDA-SN-7481-2; K. Schurer and E. Higgs, *Integrated Census Microdata (I-CeM) Names and Addresses, 1851–1911: Special Licence Access* [data collection], 2nd ed. (UK Data Service, 2022) SN: 7856, accessed 13 February 2026, http://doi.org/10.5255/UKDA-SN-7856-2.

34 For extensive I-CeM documentation, see https://www.campop.geog.cam.ac.uk/research/projects/icem/ (accessed 13 February 2026). For details on pre-processing, cleaning and standardisation of I-CeM in creating AddressGB, see J. Rhodes, 'Geo-coding Addresses in Historic British Census Data: An Open Methodology'. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 58, no. 1 (2025): 31–53, accessed February 13, 2026, https://doi.org/10.1080/01615440.2024.2431491.

35 L. Shaw-Taylor and E.A. Wrigley, 'Occupational Structure and Population Change', in *The Cambridge Economic History of Modern Britain: Volume I, 1700–1870*, ed. R. Floud, J. Humphries and P. Johnson (Cambridge University Press, 2014), 53–88.

36 Bogart et al., 'Railways, divergence, and structural change'; Dan Bogart, 'The Transport Revolution in Industrializing Britain', in *The Cambridge Economic History of Modern Britain: Volume I, 1700–1870*, ed. R. Floud, J. Humphries, P. Johnson (Cambridge University Press, 2014); X. You, D. Bogart, E. Alvarez, A.E.M. Satchell and L. Shaw-Taylor, 'Transport Development and Urban Population Change in the Age of Steam: A Market Access Approach', Campop Transport Working Papers 9, accessed 13 February 2026, https://www.campop.geog.cam.ac.uk/research/occupations/outputs/preliminary/marketaccessandsteam.pdf.

37 T. Lan and P. Longley, 'Geo-Referencing and Mapping 1901 Census Addresses for England and Wales', *ISPRS International Journal of Geo-Information* 8, no. 8 (2019): 320, accessed 13 February 2026, https://doi.org/10.3390/ijgi8080320; T. Lan and P. Longley, 'Urban Morphology and Residential Differentiation across Great Britain, 1881–1901', *Annals of the American Association of Geographers* 111, no. 6 (2021): 1796–1815, accessed 13 February 2026, https://doi.org/10.1080/24694452.2020.1859982. For alternative methods of working with historic census data at a sub-parish level, see Niall Cunningham and Ian Gregory, 'Hard to Miss, Easy to Blame? Peacelines, Interfaces and Political Deaths in Belfast during the Troubles', *Political Geography* 40 (2014): 64–78, accessed 13 February 2026, https://doi.org/10.1016/j.polgeo.2014.02.004.

38 For a full discussion of the geo-coding method, see Rhodes, 'Geo-coding Addresses'; CensusGeocoder pipeline code: https://github.com/Living-with-machines/CensusGeocoder (accessed 13 February 2026).

39 https://www.ordnancesurvey.co.uk/business-government/products/open-map-roads (accessed 13 February 2026).

40 'GB1900 Gazetteer' made available by the GB1900 Project. We acknowledge the Great Britain Historical GIS Project at the University of Portsmouth, the GB1900 partners and volunteers. https://www.pastplace.org/data/#tabgb1900 (accessed 13 February 2026).

41 J. Rhodes, 'AddressGB: Geo-coded British Census Addresses, 1851–1911', Zenodo, 6 November 2024, accessed 13 February 2026, https://doi.org/10.5281/zenodo.10473597. For the accompanying evaluation dataset, see J. Rhodes. 'AddressGB Manual Evaluation Sample'. Zenodo, 6 November 2024, accessed 13 February 2026, https://doi.org/10.5281/zenodo.13770048.

42 Rhodes, 'Geo-coding Addresses', 42.

43 Rhodes, 'Geo-coding Addresses', 44–9.

44 Rhodes, 'Geo-coding Addresses', 44–9.

45 James Johnson and Colin G. Pooley, 'The Internal Structure of the Nineteenth-Century British City: An Overview', in *The Structure of Nineteenth Century Cities*, ed. James Johnson and Colin G. Pooley (Croom Helm, 1982), 9.

46 For several key essays, see J.B. Harley, *The New Nature of Maps: Essays in the History of Cartography*, ed. Paul Laxton (Johns Hopkins University Press, 2001).

47 Hosseini et al. 'Maps of a Nation?', 286.

48 The version of MapReader used in this work was first published in Kasra Hosseini, Daniel C.S. Wilson, Kaspar Beelen and Katherine McDonough, 'MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale', *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities* (ACM: New York, 2021): 8–19, accessed 13 February 2026, https://doi.org/10.48550/arXiv.2111.15592. See also Chapter 2 in this volume and Rosie Wood, Kasra Hosseini, Kalle Westerling, Andrew Smith, Kaspar Beelen, Daniel C.S. Wilson and Katherine McDonough, 'MapReader: Open Software for the Visual Analysis of Maps', *Journal of Open Source Software* 9, no. 101 (2024): 6434, accessed 13 February 2026, https://doi.org/10.21105/joss.06434.

49 Hosseini et al., 'MapReader: A Computer Vision Pipeline': 4. It was beyond the scope of our original work using MapReader to detect tramlines on maps (since these would have required another set of manual annotations). It will be possible to produce a dataset of railways inclusive of tramways using MapReader in the future.

50 The number of patches annotated for this experiment is not prescriptive: how many patches are required for training, testing and validating a model will vary from experiment to experiment.

51 For further details and to access the model, see https://huggingface.co/Livingwithmachines/mr_resnest101e_timm_pretrain. The model is highly accurate in detecting rail infrastructure, as shown by an F1 score of 96.74 per cent. The F1 score is a combined assessment of the model's ability to correctly detect the presence of rail infrastructure (true positives) while minimising instances of rail being detected where there was none (false positives), and detect the absence of rail infrastructure (true negatives) while minimising the amount of rail infrastructure not identified as such (false negatives). For further details, see Hosseini et al., 'MapReader: A Computer Vision Pipeline'.

52 Kasra Hosseini, Daniel C.S. Wilson, Kaspar Beelen and Katherine McDonough, 'Mapreader_data_sigspatial_2022', Zenodo, October 5, 2022, accessed 13 February 2026, https://doi.org/10.5281/zenodo.7147906. For details on the dataset creation using MapReader, see https://github.com/Living-with-machines/MapReader/wiki/GeoHumanities-workshop-in-SIGSPATIAL-2022. It is important to state that this dataset is not error free: after fine-tuning a 'weak model' (e.g. it was not trained using a large number of labelled patches), we used the results to visually identify systematic errors in the predicted patch labels. We then used annotation targeted on the areas of sheets containing such errors to add to the labelled dataset. Some automatic post-processing of the predicted railspace results removed isolated patches (e.g. railspace patches with no neighbouring railspace patch within 205 metres were unclassified as railspace). 16,439 sheets produced about 30.5 million patches, of which about 1 per cent were classified as railspace.

53 M.H. Cobb, *The Railways of Great Britain: A Historical Atlas at a Scale of 1 Inch to 1 Mile* (Ian Allan Publishing, 2006). The Campop railway datasets are based largely on the Cobb atlas: J. Marti-Henneberg, M. Satchell, X. You, L. Shaw-Taylor and E. Wrigley, *1881 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852993, accessed 13 February 2026, https://doi.org/10.5255/UKDA-SN-852993; J. Marti-Henneberg, M. Satchell, X. You, L. Shaw-Taylor and E. Wrigley, *1861 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852992, accessed 13 February 2026, https://doi.org/10.5255/UKDA-SN-852992; and M. Satchell, E. Wrigley, L. Shaw-Taylor, X. You and J. Henneberg, *1851 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852991, accessed 13 February 2026, https://doi.org/10.5255/UKDA-SN-852991.

54 Coll Ardanuy et al., 'Station to Station'; https://rchs.org.uk/railway-passenger-stations-in-great-britain-a-chronology/, version 5.02 released September 2020 by the Railway and Canal Historical Society (accessed 14 September 2021).

55 We only use stations in *StopsGB* with a confidence score (a measure of the certainty that the station has been geo-coded correctly) of greater than 50 per cent.

56 *StopsGB* collapses the differences between passenger stations (e.g. urban, suburban, mainline), although we have information about which company operated trains from each station.

57 The method for calculating the railspace score differs depending on whether the street geometry is an OS Open Roads line or a GB1900 point. For OS Open Roads, we split the road into a series of points at each vertex (whenever the road changes direction), calculating the proportion of MapReader patches identified as railspace within 100 m radius of each point. The railspace score for the street is the mean of these scores. For GB1900, we simply calculate the proportion of MapReader patches identified as 'railspace' within 100 m radius of the GB1900 point.

58 B. Seebohm Rowntree, *Poverty: A Study of Town Life* (Macmillan, 1901), xi, 14; also Charles Booth, *Labour and Life of the People of London*, vol. 1, *East London* (Macmillan, 1889), 60.

59 The schema for Scotland lacks two transitional categories found in the England and Wales schema. To harmonise the schemas for Great Britain, we have classified 'Urban transition' parishes as 'Urban' and 'Rural transition' parishes as 'Rural'. H. Smith, R. Bennett and D. Radicic, 'Towns in Victorian England and Wales: A New Classification', *Urban History* 45, no. 4 (2018): 568–94, accessed 13 February 2026, https://doi.org/10.1017/S0963926818000020; Harry Smith and Robert J. Bennett, 'Urban-Rural Classification Using Census Data, 1851–1911', Drivers of Entrepreneurship and Small Business, Working Paper 6 in Working paper series from ESRC project ES/M010953 (2017), accessed 13 February 2026, https://doi.org/10.17863/CAM.15763.

60 Jackson, *Semi-Detached London*, 21.

61 Unfortunately, I-CeM's servant variable does not differentiate between different types of servants, for example, those working in hospitality premises or servants living and working on farms.

62 For details of the I-CeM occupational codings used to develop our clusters see Table 3.6.

63 Walking remained a common mode of transport for all workers, and especially those clerks who could live in the inner suburbs, of whose number 22,000 lived in Camberwell by 1901, see Dyos and Reeder, 'Slums and Suburbs', 368, 371.

64 Dyos, *Victorian Suburb*, 69–74, 109–13.

65 George and Weedon Grossmith, *The Diary of a Nobody* (Penguin, [1892] 1985), chapter 1.

66 Simon Abernethy, 'Class, Gender and Commuting in Greater London, 1880–1940', (Unpublished PhD thesis, University of Cambridge, 2015), chapter 6, 'On the Buses'.

67 The Manchester Ship Canal, which served Ordsall's Salford Docks complex, only opened in 1894.

68 Robert Roberts, *The Classic Slum: Salford Life in the First Quarter of the Century* (Manchester University Press, 1971).

69 Roberts, *The Classic Slum*, 3.

70 Lady Bell, *At the Works: A Study of a Manufacturing Town* (T. Nelson & Sons, 1907), 14, 43 and 215, accessed 13 February 2026, https://archive.org/details/atworksstudyofma00bellrich.

71 Places (a mixture of parishes and sub-parish units) derived from I-CeM's ParID variable. We excluded locations with fewer than the median number (ten) geo-coded streets per parish. To aid graphic representation, we also filtered out remote rural locations where the average distance to a station was over 15 km.

72 LSE Library, BOOTH/B/349, 1897, George H. Duckworth's Notebook: Police and Publicans District 10 [Bethnal Green East], District 15 [South West Islington], District 17 [Upper Holloway], 179–81.

73 LSE Library, BOOTH/B/348, 1897, George H. Duckworth's Notebook: Police and Publicans District 14 [West Hackney and South East Islington], District 15 [South West Islington], District 16 [Highbury, Stoke Newington, Stamford Hill], 235–9.

74 LSE Library, BOOTH/B/349, 1897, George H. Duckworth's Notebook: Police and Publicans District 10 [Bethnal Green East], District 15 [South West Islington], District 17 [Upper Holloway]], 21–3.

75 Jerry White, *The Worst Street in North London: Campbell Bunk, Islington, between the Wars* (Routledge & Kegan Paul, 1986), 11–14.

76 Dyos, *Victorian Suburb*, 112.

77 LSE Library, BOOTH/B/365, 1899, George H. Duckworth's Notebook: Police District 32 [Trinity Newington and St Mary Bermondsey], District 33 [St James Bermondsey and Rotherhithe], District 34 [Lambeth and Kennington], District 35 [Kennington (2nd) and Brixton], District 41 [St Peter Walworth and St Mary Newington], District 42 [St George Camberwell], District 45 [Deptford], 103–5.

78 White, *Worst Street*, 8–70.

79 *Northern Weekly Gazette*, 3 August 1895, accessed 13 February 2026, https://www.britishnewspaperarchive.co.uk/viewer/bl/0003074/18950803/149/0007.

80 Roberts, *Classic Slum*, 3.

81 Dennis, *Industrial Cities*, 110–32; Geddes, *Cities*, 24, 34–45.

82 A publicly available code repository available here: https://github.com/Living-with-machines/beyondthetracks. Data is available here: https://doi.org/10.5281/zenodo.15175157.

Chapter 4

Links in the chain: workers in the British bicycle boom, as seen from the England and Wales census, 1881–1911

Guy Solomon, Joshua Rhodes, Jon Lawrence, Emma Griffin, Timothy Hobson, Sherman Lo and Daniel C.S. Wilson[1]

Introduction

On the Living with Machines project, most of our research using historic census data focused on spatial patterns that could be explored using anonymised, individual-level data. In 'Beyond the tracks' (Chapter 3) we developed new methods for aggregating individual and household data at street level to explore the social and environmental impact of the rail network in novel ways. Working with *un*-anonymised, individual-level census data is more difficult, with privacy and copyright requiring higher levels of data security and influencing what can be put into the public domain. But the rewards are also considerable. Perhaps most importantly, working with individual-level data makes it possible to trace people over time. Because the census only happened every ten years, people could slip through its net or be rendered otherwise untraceable; nonetheless, historians have long made use of the population census as a powerful tool for recreating the broad contours of lives that otherwise left few archival traces.[2] Family historians have turbo-charged this method, using it to reconstruct extensive personal genealogies that were previously the preserve of a narrow social elite.[3] As part of the Living with Machines project, we sought to develop new ways to use computational methods to scale-up such uses of the census. Our goal was to enable researchers to reconstruct life-course social

change at the level of the group rather than just the isolated individual. In short, we sought to develop a new, reproducible (and refinable) method for automatically linking large numbers of people between census years. The challenge was to come up with a method that would deliver high-quality matches at a scale sufficient to allow us to reconstruct the social experience of a whole social group rather than just a small cluster of individuals. The resulting method for automatic census linking we outline in this chapter represents a tool that can deepen our understanding of the human story behind major processes of social and economic change in Victorian and Edwardian Britain from internal migration and marriage practices to occupational change.

As one application of this method, this chapter focuses on labour recruitment to one of the most dynamic and transformative new industries of the late nineteenth century: the cycle trade.[4] Although companies had been making different types of cycle on a significant scale in Britain from the late 1860s, the industry really took off in the later 1880s, driven mainly by a series of British technical innovations (notably the 'Rover' safety frame and the Dunlop pneumatic tyre) that improved both the ease and comfort of cycling, transforming it from a hobbyist's eccentric enthusiasm into a mass leisure pursuit for old and young, men and women.[5] Initially a luxury item, over the following decades the bicycle became a machine that changed everyday life for the masses. In 1881 there were barely a thousand workers employed in cycle making across Britain, ten years later this had risen to approximately 10,000, and at the height of the 'cycle mania' in 1897 the industry employed approximately 50,000 workers, with the English West Midlands, and especially the cities of Coventry and Birmingham, at the epicentre of the trade.[6] We know a great deal about the pioneers of this revolution. Many had their roots in other branches of the metal trade, or in consumer-facing manufacturing: some of the most famous pioneers of the British cycle trade, such as James Starley, William Hillman and George Singer began life on the shopfloor as skilled mechanics.[7] It is with good reason, therefore, that the cycle boom of the 1880s and 1890s has been used to refute claims that 'gentrification' had suffocated the 'industrial spirit' of late Victorian Britain.[8]

Cycle manufacturing grew up wherever there was an existing concentration of metalworking and machine-making expertise, most notably in the English Midlands, but also in other towns and cities where these trades were well established. It is generally argued that the new industry

flourished where it could secure a good supply of highly skilled metalworkers and machinists.[9] But did such workers form the mainstay of the new industry's labour force? Or, did the cycle industry suck in labour from other trades, and from further afield, as it expanded in the late nineteenth century? Did the cycle boom create new opportunities for workers to move into the metal trades, picking up skills on the job? If so, did new technologies such as cycle making act as vehicles for upskilling previously excluded demographic groups?

Applying our new census-linking method makes it possible to address such questions for the first time on an industry-wide scale. It allows us to identify who the new cycle makers were: what they had done beforehand, where they had lived, and what they went on to do (if they left the industry). In short, it unlocks the story of the sudden emergence of the cycle industry by focusing on changes on the shopfloor rather than the boardroom. This shift in perspective allows us to understand the social effects of technological change in new ways. We pay special attention to two cases: Birmingham and Coventry. These were the most important centres of cycle production during the boom years of the 1890s, but they provide contrasting stories. In the later nineteenth century, Birmingham was establishing itself as the metropolitan centre of the wider West Midlands region, with a diverse economy anchored in the metal trades but within which the cycle industry played only a minor role. Coventry was a much smaller city (approximately one-eighth Birmingham's size in 1901), but its cycle industry was larger. The rise of cycle manufacturing transformed Coventry's economic and demographic fortunes, laying the foundations for its emergence as a major centre of light engineering in the twentieth century.[10] The relative importance of the cycle trade meant it would have different social impacts in these two cities.

Geography, innovation and labour in Britain's bicycle boom

Britain's industrialisation not only occurred early, it was also highly regionalised, with different parts of the country coming to dominate whole sections of industrial production: notably Lancashire with cotton, Yorkshire with woollens, the East Midlands with lace-making, and the West Midlands with light metalworking.[11] Although never holding a monopoly over production, the regions' pre-eminence often led to sub-regional specialisation.

Just as specific towns came to dominate different branches of the cotton and woollen trades, a similar pattern for metals developed in the West Midlands.[12] The small township of Wednesfield, outside Wolverhampton, was said to have a near monopoly on England's animal trap industry in the late nineteenth century, while nearby Willenhall dominated the country's lock and key making trades.[13] Although larger centres such as Birmingham and Coventry had more diverse economies, different branches of the metalworking trade nonetheless remained pre-eminent. Birmingham had jewellery-making and gun-making, as well as a wide range of general metalware industries; Coventry had watch and clock making, as well as textile machine-making, but it was also a historic centre of silk weaving and ribbon making.[14] By the late nineteenth century, large firms were beginning to emerge in many of these trades, including BSA for armaments, Nettlefolds for hardware, and Chubb for locks and safes, but the backbone of manufacturing across the West Midlands remained the small workshop rather than the mass-production factory. Indeed, even those companies that invested heavily in factory-based production often continued to make extensive use of small-scale sub-contractors for specialist components or for finishing processes.[15]

It was out of this milieu that Britain's formidable cycle industry emerged between the 1860s and 1880s. A myriad of technical innovations contributed to the cycle-making revolution. New machine tools, initially developed for the US armaments industry, stimulated innovative production processes across the engineering sector internationally, including in textile machine-making, where the sewing machine proved especially revolutionary.[16] Many of the early cycle companies, including major Coventry firms such as Starley Brothers, Hillman & Herbert, and the Coventry Machinists Company, grew directly out of the local sewing machine industry.[17] This pattern of diversification was also important in other English cities including London, Birmingham and Leicester, but it was not the only route into the new industry. In Leicester and Nottingham, companies specialising in textile machinery associated with the local hosiery and lace trades diversified into cycle making, while in Birmingham and London armaments companies did the same.[18]

But the West Midlands emerged as the epicentre of the new industry because of its concentration of expertise in diverse aspects of metalworking, including specialist tube-making, which proved vital to the development of lighter, stronger cycles and the development of ball bearings, which

helped reduce friction between the bicycle's moving metal parts.[19] Taking the Coventry cycle trade as his case study, Norcliffe frames this process as the 'geographical construction of technology', marked by networks of knowledge at regional level, driving both innovation and its commercial exploitation.[20] Other regional specialisms, such as saddlery and fine instrument-making, fed the development of ancillary cycle component industries across the West Midlands. Indeed, the regional component industry grew in parallel with the major cycle manufacturing companies such as Premier, Humber, Swift and Rudge-Whitworth (all Coventry-based firms).[21] Both adopted mass-production methods to meet burgeoning demand and growing competition from America, but the ready availability of machined components meant that the small-scale, bespoke bicycle maker nonetheless remained central to the British cycle industry throughout this period.[22] Those big-name Coventry cycle brands might all employ a thousand (or more) workers at the height of the 1890s boom, but nationwide the *average* number of employees per bicycle manufacturer barely changed across the boom years of the 1890s (it was forty-six in 1890, forty-two in 1895, fifty in 1896 and forty-three in 1897).[23]

In the late 1890s, many British bicycle companies went to the wall, or drastically scaled back production, as American firms, facing a crisis of overproduction, dumped cheap bikes on Britain's open market.[24] But those that survived the shock, or set up in its wake, continued to innovate. In these years, British companies were quick to adopt improvements to the ease and comfort of cycling such as the freewheel, rim brakes, shock absorbers and cushioned seats, while Raleigh of Nottingham pioneered the first reliable gear system: the three-speed Sturmey-Archer hub.[25] At the same time, many cycle firms, especially in the West Midland metalworking centres of Coventry, Birmingham and Wolverhampton, began to experiment with the production first of motorcycles and then cars.[26] Indeed, this diversification was a major reason why the rapid growth in cycle employment slowed after 1901. Many of the names of late Victorian cycle manufacture became famous motorbike and car marques, including BSA, Triumph, Rover and Humber. Even so, on the eve of the First World War, British cycle firms still dominated the world's export markets, and the bicycle had become the staple means of transport for working people across the industrialised world.[27]

Although the concentration of metalwork expertise in the West Midlands played a role in the emergence of its cycling industry, this did

not necessarily mean its workers had a background in metalworking. The metalworking sector was growing rapidly during this period, and cycle firms may have needed to draw in labour from a broader pool to meet their needs. Potential sources for this labour included young adults with little prior work experience, established workers from other sectors, and potentially, as some historians have claimed, from large-scale migration to boom towns such as Coventry.[28] Understanding who the bicycle makers were is key to understanding what Norcliffe's 'geographical construction of technology' meant in practice: our census-linking method makes this possible for the first time.

Data and methodology

We locate workers employed in the bicycle industry using individual-level census data, which form part of a larger dataset, I-CeM (Integrated Census Microdata), comprised of standardised and coded variables derived from digital transcriptions of census returns between 1851 and 1911.[29] The census has been a rich source for historians investigating the nature and extent of occupational change: by 1851, British census returns typically included name, street address, parish of residence and of birth, county and country of birth, age, sex, relationship to others in the household, and occupation. From this date, therefore, it becomes possible to identify named individuals in the census and to chart occupational change over time.

We extract two occupational groupings from the census – workers in bicycle manufacturing (which we henceforth call 'bicycle makers') and a composite group of light metalworkers that excludes the cycle industry (details on the construction of both groups are provided in the Appendix). The purpose of comparing bicycle makers with a group of composite metalworkers is not to position this as an apples-to-apples comparison – the latter contains a broad array of industries, each with their own technological developments and economic conditions. Rather, we use the composite metalworkers as an illustration of the broader trajectory of comparable metalworking trades, which allows us to understand how far bicycle makers diverged from the norm. Both the bicycle maker and composite metalworker occupational groupings are drawn from the 'Occode' variable in I-CeM – a classification scheme of 797 occupational categories.[30]

We start our analysis in 1881, since the England and Wales 1871 census is not included in I-CeM. Although this truncates our history of 'feeder' industries (such as gun-making or sewing machine manufacturing), we still capture most bicycle makers, since there were very few bicycle manufacturers established before 1881. The importance of exploring the shift into new industries by 1911, especially motor vehicles, requires confining analysis to England and Wales given the absence of Scottish data for 1911.[31] In 1881, we identify 949 bicycle workers in England and Wales (891 men, 58 women). In 1891, there were 9,635 (8,777 men and 457 women); in 1901, there were 27,393 (23,965 men, 2,456 women); in 1911, 25,138 (21,983 men, 3,153 women).[32]

Throughout the period, there were many more composite metalworkers than bicycle makers: around thirty-five times as many at the national level, even at the high point of cycling employment in 1901. However, one element which is not immediately evident from total employment figures is the spatial distribution of each occupational grouping. As discussed above, bicycle manufacturing was heavily concentrated in the Midlands – and Warwickshire in particular (Figure 4.1). It is for this reason that we select Coventry and Birmingham (including Smethwick) for particular focus, as locations in which bicycle manufacturers were concentrated, in addition to undertaking analysis at a national level. By contrast, our composite metalworkers were far more spatially distributed, though with especially heavy concentrations in Lancashire, the Midlands, Yorkshire and London (Figure 4.2). In 1901, the composite metalworkers grouping outnumbered bicycle makers six to one in the latter's heartland of Warwickshire – a sizable number for certain, but far less than the dramatic difference nationally.

Having identified bicycle makers and composite metalworkers, our next task is to locate them in the preceding and subsequent censuses to reconstruct their occupational trajectories and migration histories. To do this, we have developed an automated Python pipeline to link individuals between censuses. Automated record linkage is a well-established field, routinely used today to link individuals across databases, including patient records, tax records and census data. Historians have also used record linkage in similar ways, both to link individuals across different sources and across the same source base over time.[34] Record linkage has been used in relation to American census data, for example, in The Census Tree



Figure 4.1a: The distribution of bicycle makers, by census year: a) 1881, b) 1891, c) 1901, d) 1911. Counties not directly matched to individuals for 1911 given in grey. Credit: Guy Solomon.[33]

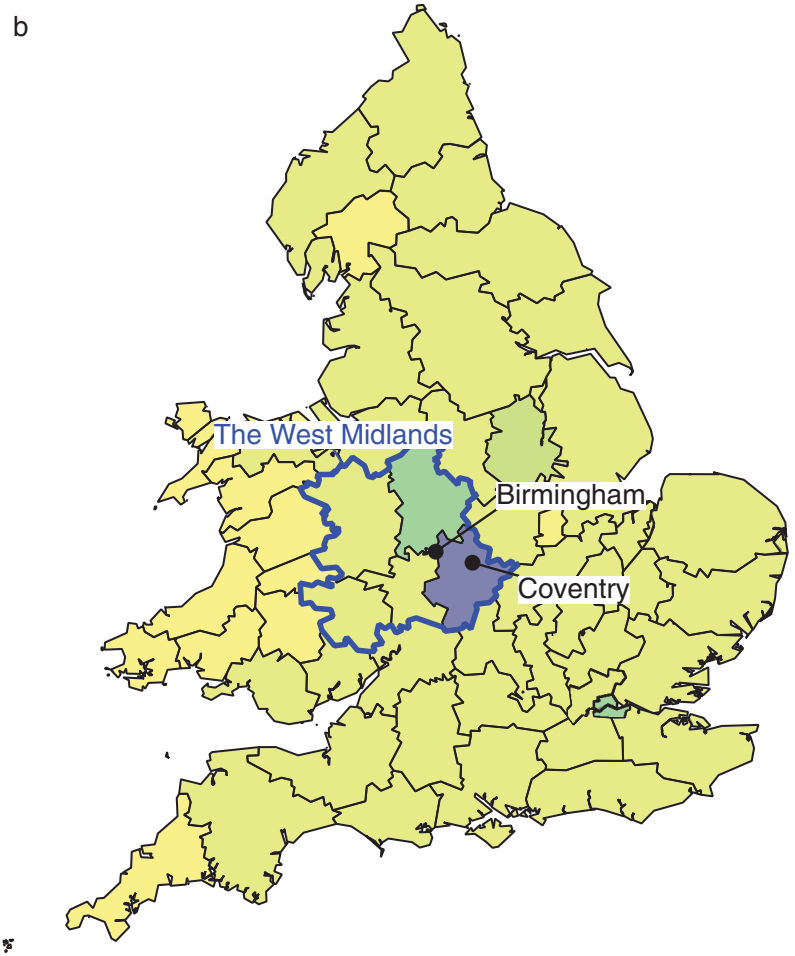


Figure 4.1b: (cont.).

project, which links 700 million US census entries between 1840 and 1950, through a combination of machine learning and manual links generated by users of a major genealogy website.[35] However, examples of the automated linking of historic British census records remain comparatively rare, owing to access restrictions on individual names data, which is an essential prerequisite for linkage.[36]

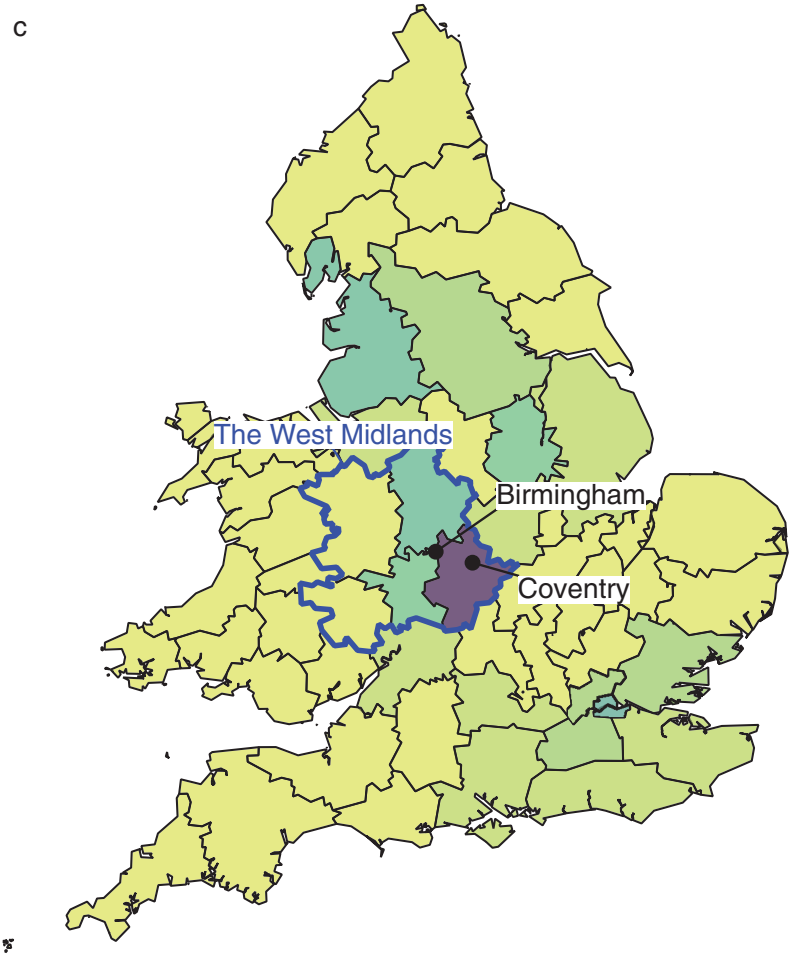


Figure 4.1c: (cont.).

Our approach to matching individuals across British censuses uses methods drawn from digital humanities and data science. We use the combination of key data recorded by the census – name, gender, age and place of birth – as a unique personal fingerprint, without using mutable characteristics, such as occupations, given that our central focus here is occupational change.

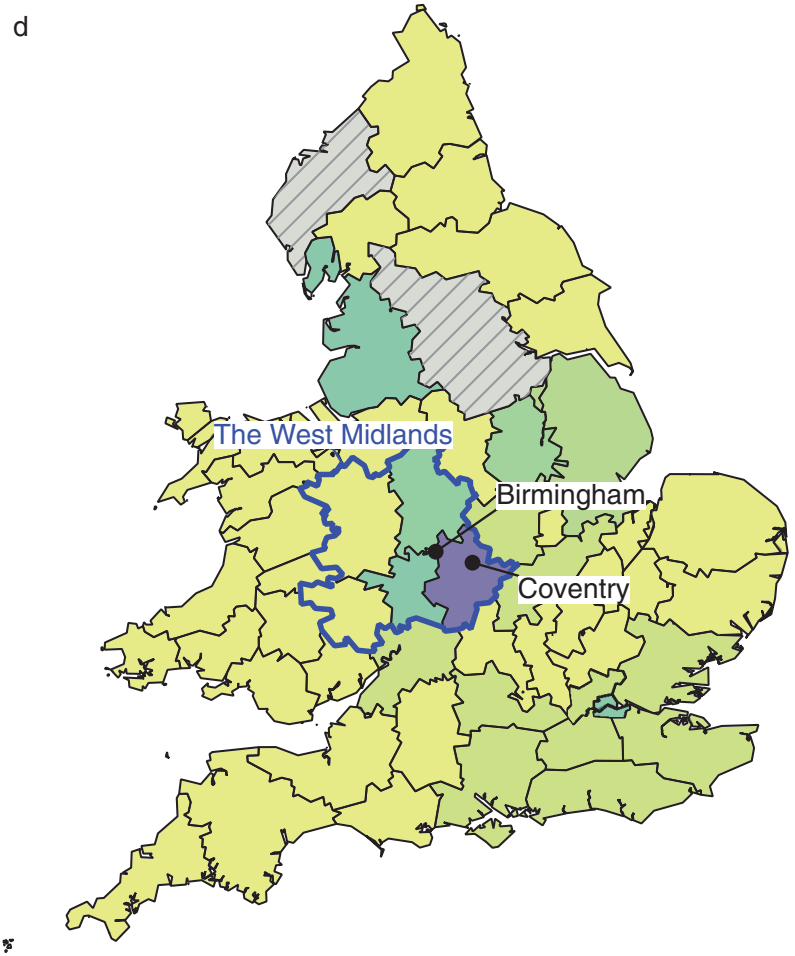


Figure 4.1d: (cont.).

Table 4.1 outlines how we match each of these criteria to determine if two persons in different censuses are likely the same individual. The matching process is split across three 'scenarios'. Each scenario is run in both directions between the two census years to ensure that the best candidate for a match in one direction (e.g. from 1881 to 1891) is also the best candidate in the reverse (e.g. 1891 to 1881). Scenario 1 finds individuals whose

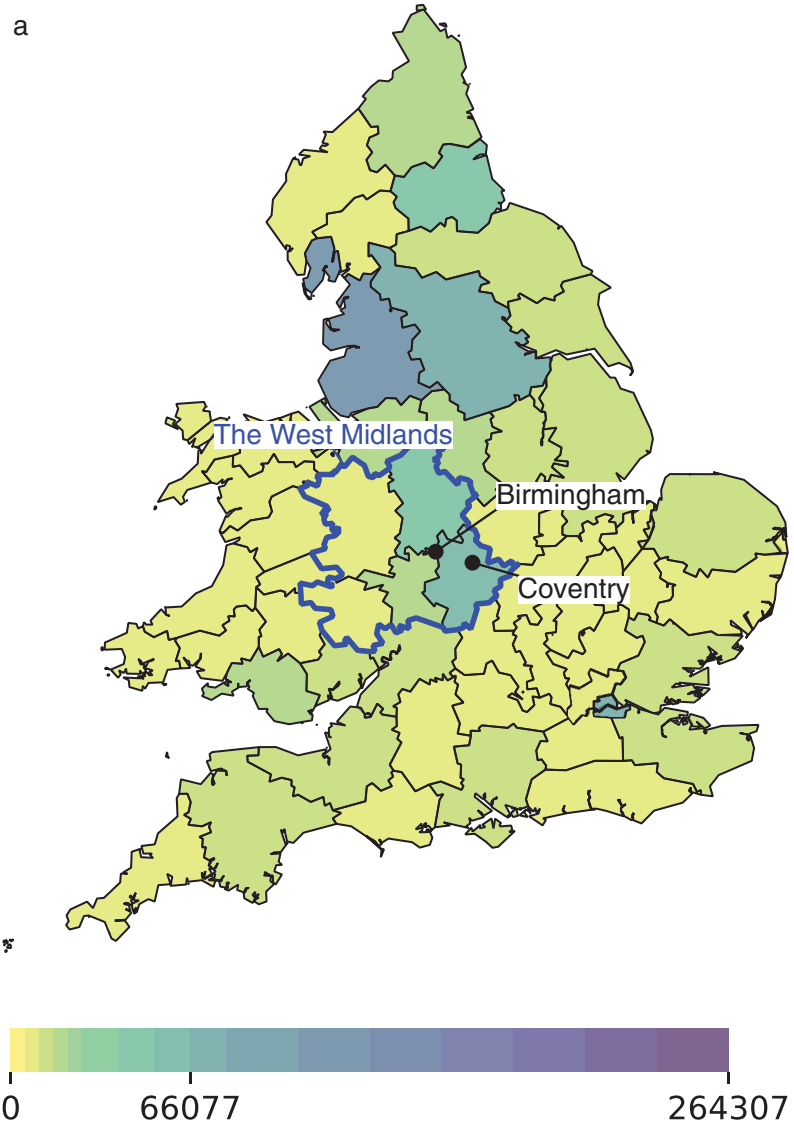


Figure 4.2a: The distribution of composite metalworkers, by census year: a) 1881, b) 1891, c) 1901, d) 1911. Credit: Guy Solomon.

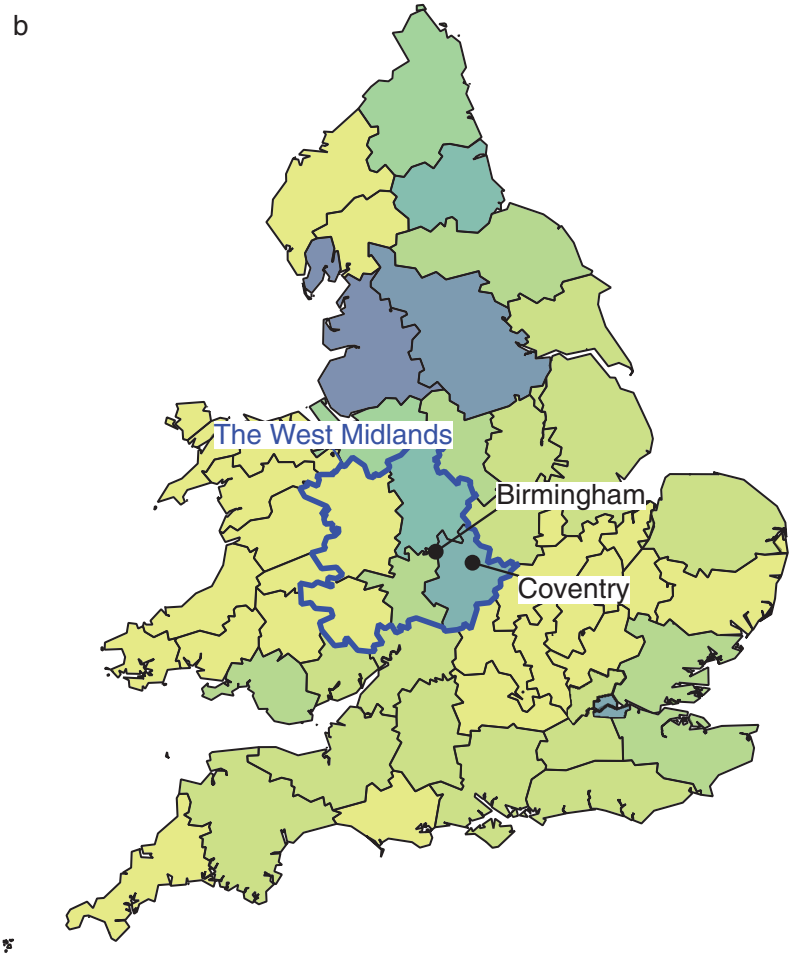


Figure 4.2b: (cont.).

personal information matches exactly (except age, where we allow some variation due to the variance in how people reported their ages). This is the 'strictest' matching scenario, since almost all information must match exactly. Nevertheless, it is also responsible for around 80 per cent of our matches. Individuals successfully matched in Scenario 1 are then removed from the census pool, before running Scenario 2 on the remaining individuals. Scenario 2 relaxes the requirements for exact matches, allowing for

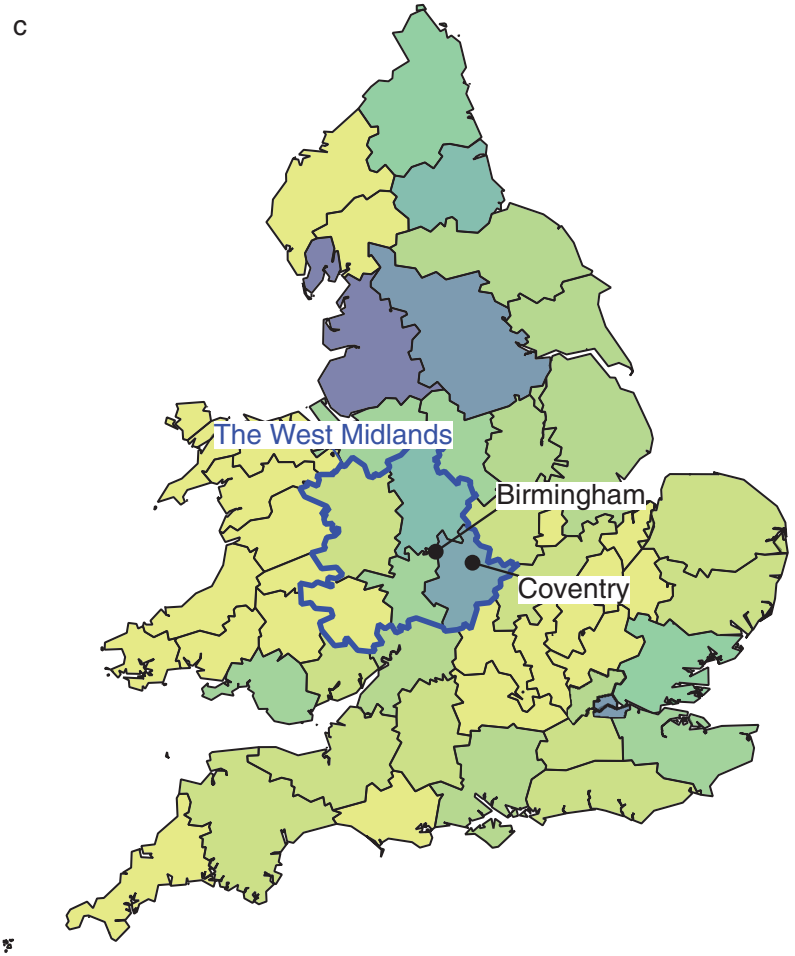


Figure 4.2c: (cont.).

some variation in spelling of names and places of birth.[37] Entries meeting these matching criteria are also removed from the census pool prior to running the next scenario. There are instances where more than one record would meet the requirements for a successful match in either Scenario 1 or 2. In this case, the entire *nominal linkage* method is applied to the household members of potential matches, and successful matches used as a tiebreaker. Should this still fail to distinguish between the two records,

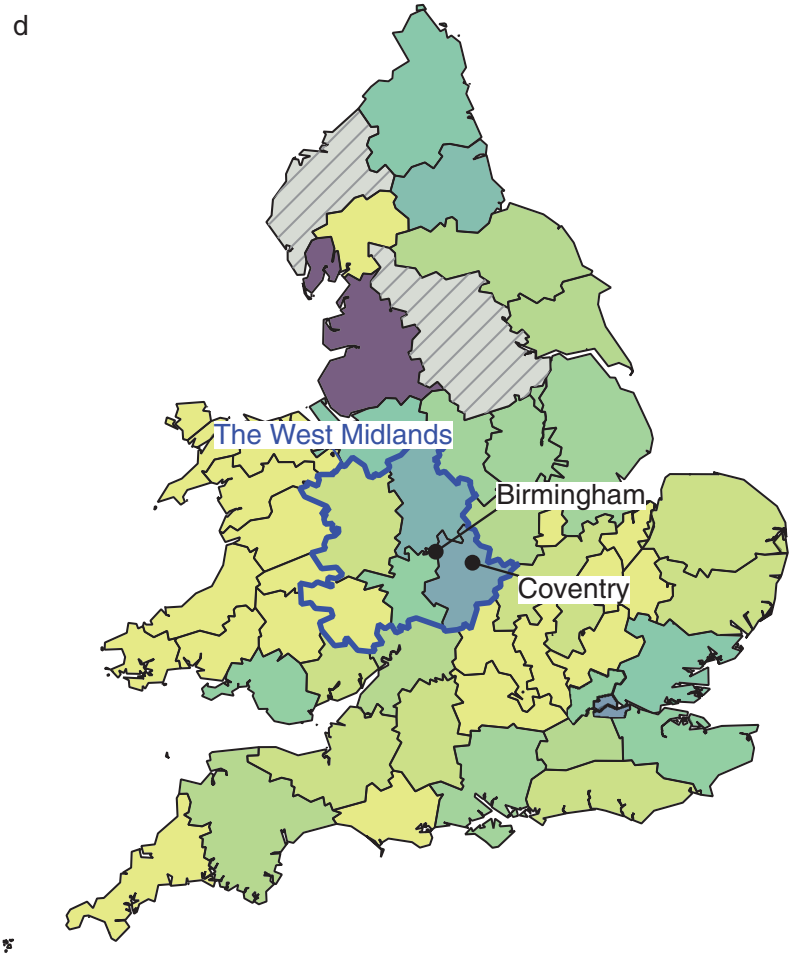


Figure 4.2d: (cont.).

neither is accepted and the individual is returned to the candidate pool for the next scenario.

Scenario 3 is aimed specifically at identifying women whose surname may have changed upon marriage between two consecutive census dates. The comparison of surnames is dropped entirely. Instead, the presence of co-resident individuals in the households of potential matches is used to determine whether or not they should be linked. This implementation

Table 4.1: Nominal linkage scenarios

Criteria	Scenario 1 (individual)	Scenario 2 (individual)	Scenario 3 (household)
Forename	Exact	Exact or high similarity	Exact or high similarity
Surname	Exact	Exact or high similarity	N/A
Age	+/− 2 years from expected age	+/− 2 years from expected age	+/− 2 years from expected age
Gender	Exact (matches from Male or Female to 'Unknown' permitted)	Exact (matches from Male or Female to 'Unknown' permitted)	Exact (female only, to 'Unknown' permitted)
Place of birth	Exact	Exact	Exact

of household matching to account for women's marital name changes is important for the wider application of our linking method to other groups of people in the census. However, since the majority of metalworkers (including those employed in bicycle making) were male, Scenario 3 contributes relatively few matches in the present application.

We evaluate the accuracy of our record linkage with a manual examination of a randomly selected sample of the proposed matches. Rather than asking the question 'is this an accurate match?' (which is not a question capable of definitive answer) each expert evaluator was asked 'is this a reasonable match, given all contextual information?' On this basis, 95 per cent of the matches proposed by the algorithm were deemed to be reasonable.[38] Such evaluation offers no insight regarding false negatives, where the method fails to create a match where there should be one – instead, we consider this through the matching rates. The *nominal linkage* implementation adopted here deliberately prioritises high confidence matches over maximising the number of matches created. Although this already provides a much larger volume of data

than has typically been obtainable via manual linking, it is still necessary to ensure a large enough linked sample that any conclusions are robust to small sample effects. Figure 4.3 and Figure 4.4 demonstrate the number of linked individuals, relative to the number of people employed in each occupational group. We do not consider bicycle makers in 1881 as a group for analysis, owing to their very small number – although the overall number of people employed in this occupation is shown Figure 4.3, in order to give a sense of the growth of the industry between 1881 and 1891.[39]

As Figure 4.3 demonstrates, although the number of bicycle makers varies by census, approximately one-third of these individuals are linked (although those linked forwards to the next census are not necessarily the same individuals as those linked backwards to the previous census – 58 per cent of our 1891 bicycle makers linked backwards to 1881 were also linked forwards to 1901, and 48 per cent of 1901 bicycle makers linked backwards to 1891 were also linked forwards 1911). The overall matching rate for backwards matching 1901 bicycle makers to 1891 is slightly higher, at around 40 per cent, but slightly lower for forward matching 1901 bicycle makers to 1911, at around 30 per cent. Matching rates are highest in Coventry.

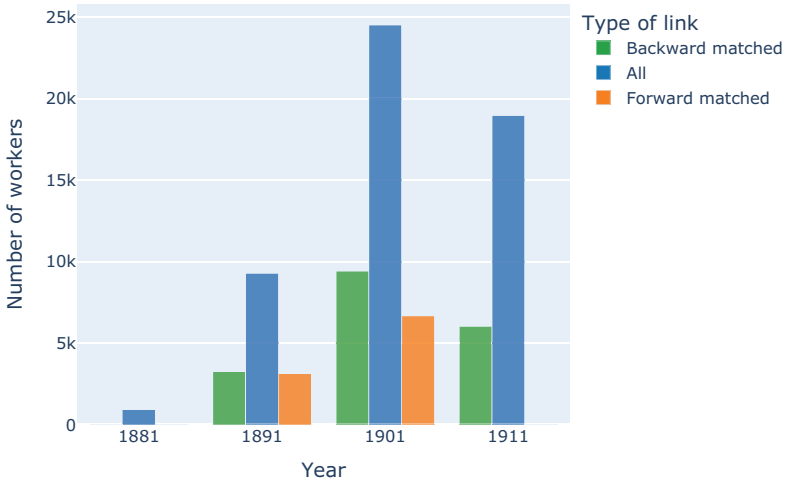


Figure 4.3: Linked bicycle makers relative to all workers in this occupation, 1881 to 1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

In Birmingham, the matching rate is lower, but the relative size of the industry – comparable to Coventry – means it still provides a sufficient number of matches for analysis. In other urban centres associated to a lesser extent with the bicycle industry (such as Wolverhampton, Nottingham, London and Liverpool), the sample of linked individuals is either too small as a proportion of the total employed in the occupation to have confidence in the representativeness of any conclusions drawn on their basis, or total numbers recorded as being employed in the trade are too low to facilitate a meaningful analysis. For our composite metalworkers classification, the picture is similar, albeit with slightly lower matching rates (Figure 4.4). Proportionally, the lowest linking rate is forwards from 1881, but given the primary role of these metalworkers in our analysis is as a control group, and that we do not undertake this exercise for bicycle makers, we can reasonably exclude these links from further discussion. Approximately 30 per cent of workers are linked both forwards and backwards from 1891, and also forwards from 1901 to 1911. Linking 1901 backwards to 1891 yields a higher matching rate, at over 35 per cent. This is also the case for linking 1911 workers backwards to 1901 (as for bicycle makers, no matches forwards from 1911 can be sought at this stage).

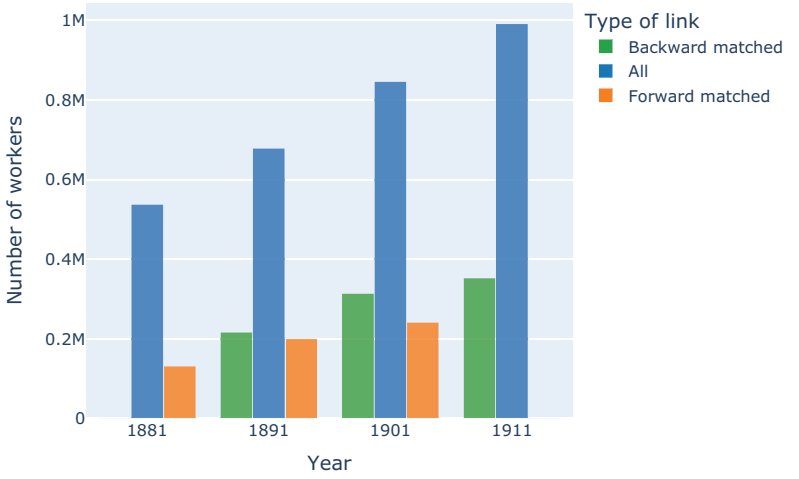


Figure 4.4: Linked composite metalworkers relative to all workers in this grouping, 1881 to 1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

Analysis: occupational origins (where did they come from?)

For a boom industry such as bicycle manufacturing, one of the major questions is where it found its labour. Reconstructing the characteristics of these workers and what they had been doing ten years earlier has the potential to shed important light on the nature of the work in the bicycle industry and how new technology was adopted. Firms which pivoted their production towards bicycle manufacturing were often part of a longer running process of technological convergence in metalworking which saw companies respond to technical innovations by shifting production to new sectors, for instance from the manufacture of small firearms in the first half of the nineteenth century, to sewing machines in the mid-century, before switching to bicycles and (eventually) motor vehicles.[40] But changes occurring at the level of the firm do not always tell us about changes at the level of individual workers. By observing the movement of bicycle makers into and out of the industry we can begin to understand the extent to which its success depended upon the regional concentration of metalworking skills.

This section therefore focuses on two questions: First, what were people doing before they became bicycle makers? Second, did recruits to bicycle making demonstrate a distinctive pattern from the wider metalworking population? Both questions, in their own ways, relate to the development and cultivation of skills in the workforce. In the case of the former, if the 'technological convergence' which facilitated the development of the bicycle industry was reliant on particular skills within the working population (in addition to the infrastructural and organisational influence of firms and entrepreneurs), then this should be evident in the people employed in these roles. The latter addresses a similar but distinct theme: the extent to which the bicycle manufacturing industry was uniquely attractive to those people positioned to enter the trade, or whether it was one of several potential alternatives. In summary, we find that during its main period of growth the bicycle industry largely recruited 'new' labour fresh into the workforce (with a smaller, but notable proportion retained or drawn from related industries), and while there is a pattern distinct from metalworkers as a whole, bicycle makers do not seem to have formed a distinct working population.

We explore the change in occupations between censuses visually using Sankey plots, such as Figure 4.5. It is worth explaining how we construct these and how they should be interpreted, as they feature prominently

throughout the remainder of the chapter. The plot shows the number of individuals who move between different occupations between two census years. The size of the connecting line (henceforth discussed as a 'flow') indicates the number of individuals. In order to examine key trends, we explore this at a sectoral level – however, given that the literature suggests a strong relationship between bicycle making and other metalworking occupations, we also provide a breakdown of the ten most common occupations within the Manufacturing sector.[41]

At a national level, the rapid growth period between 1881 and 1891 is not a story of transition for workers employed in the precursor industries to bicycle manufacturing. Rather than recruiting from those with pre-existing metalworking skillsets, the major source of recruitment for the bicycle industry was workers not previously in the workforce at all. As Figure 4.5a demonstrates, of the approximately 3,250 bicycle makers in 1891 which can be matched to the 1881 census, around 1,700 of them (or just over 50 per cent) were previously classified as 'Scholars, etc'; and a further 300 (9 per cent) possess no occupational information but were also mostly too young to have been employed.

There is some evidence of recruitment from other sectors, as one would expect, such as Manufacturing. The Manufacturing sector accounted for around 20 per cent of overall recruitment to the cycle trade, with around 40 per cent of this group coming from occupations associated with the high-level metal working skills considered essential to the industry's local take-off, such as engine- and machine-fitting and turning, as well as machinists and machine workers. Without these, the trade might not have flourished across the West Midlands, but such workers constituted a small fraction – less than ten per cent – of the total workforce. The next largest flows come from the Maker-Dealer and Construction sectors, but these each account for less than five per cent of overall recruitment. In short, unlike the firms they worked for, bicycle makers with traceable employment histories show little evidence of a mass transition from one trade to another.

Despite the rapid expansion of the industry over the 1890s, the picture is very similar ten years later (see Figure 4.5b). 'Scholars, etc' and those with no occupational information (again, the latter being largely too young to have had an occupation in the 1891 census) are still the primary source of recruitment, accounting for 50 per cent of all bicycle manufacturers in 1901. The Manufacturing sector – still led by 'engine and machine fitting/

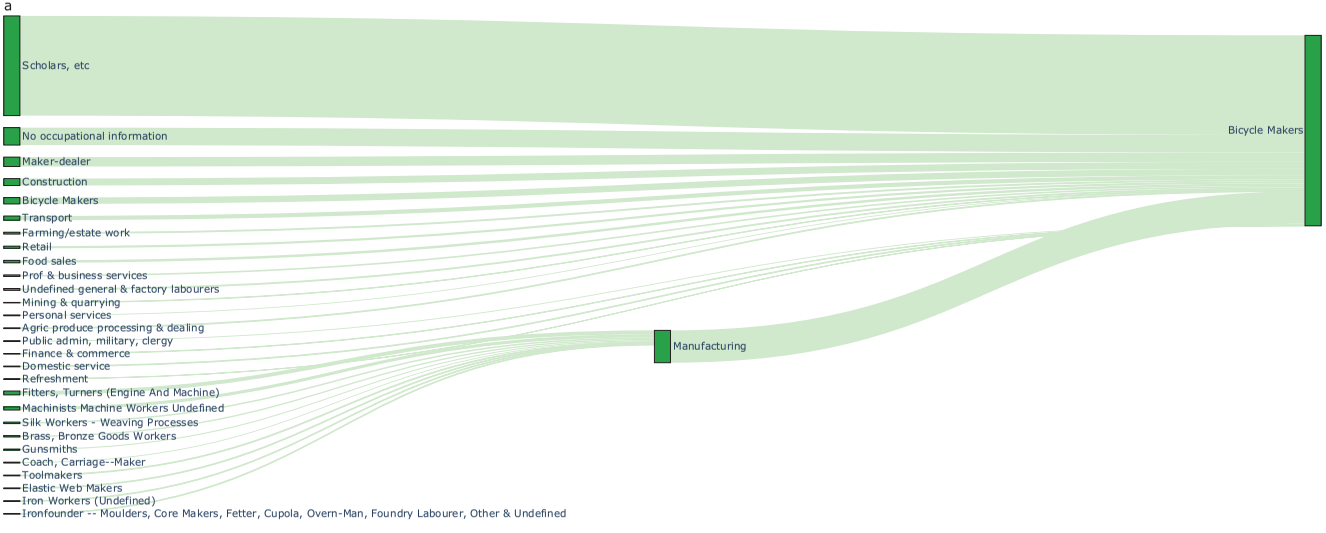


Figure 4.5a: Bicycle makers linked to the previous census: a) 1891 to 1881, b) 1901 to 1891, c) 1911 to 1901. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.
Note: Original spellings of overn-man (oven-man) and fetter (fettler) are retained.

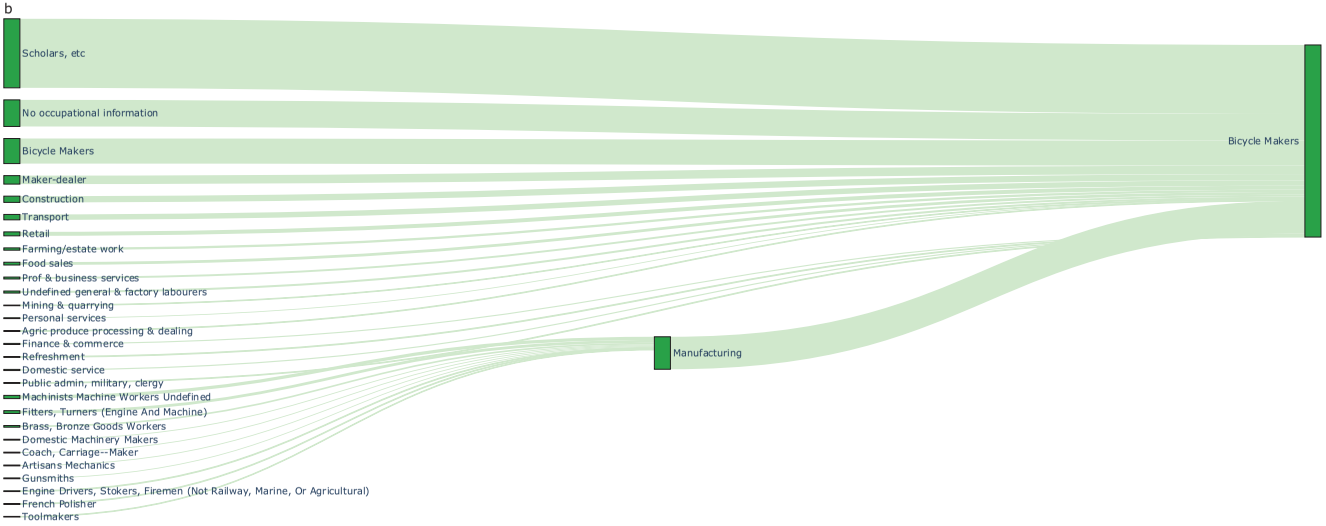


Figure 4.5b: (cont.).

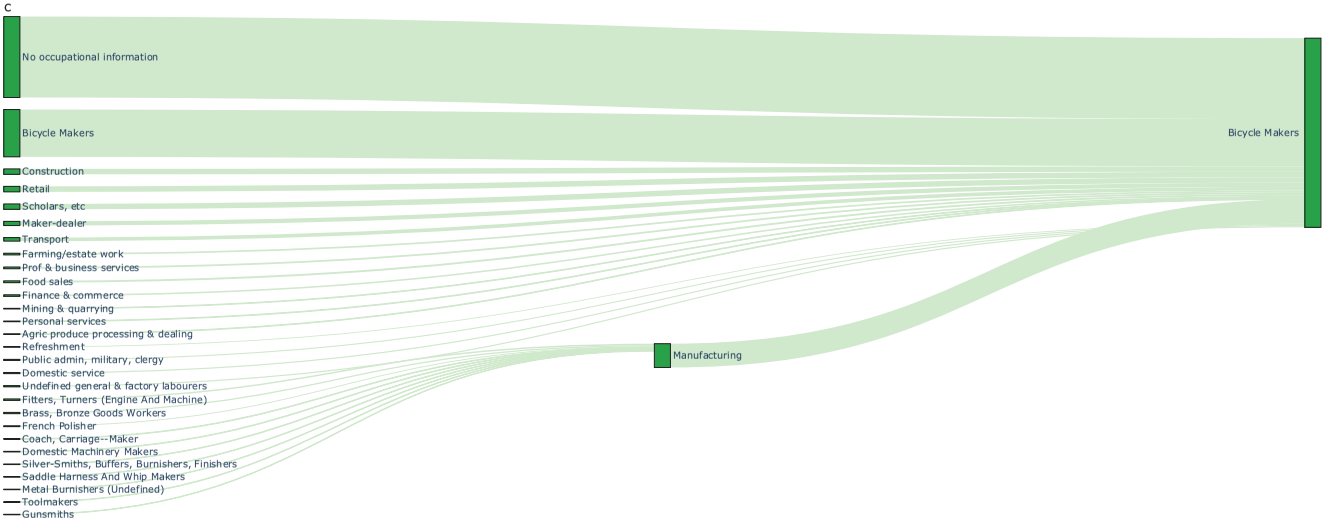


Figure 4.5c: (cont.).

turning' and 'machinists/machine workers' – again accounts for just under 20 per cent of recruitment. The largest change is the retention of workers previously employed in bicycle making, which rises to around 15 per cent. This makes sense, given the increased size of the trade relative to 1881, but nonetheless shows that the majority of individuals working in bicycle making were recruited as 'new' labour with little or no prior work experience.

Despite the slowing of the bicycle industry during the early twentieth century, Figure 4.5c also demonstrates considerable continuity for 1911. The disappearance of the 'Scholars, etc' sector appears to relate to changes in classification practice since it is largely offset by the rise in those for which we have no prior occupational information. Combined, these categories see a slight fall in their proportion of recruitment (down to a little over 45 per cent), which may indicate a slowdown in recruitment related to the industry's overall contraction, but this remains very close to the figures recorded in 1891 and 1901. The Edwardian cycle industry continued to offer employment opportunities for young workers. Recruitment from the broader Manufacturing sector similarly sees a fall of a few percentage points, but remains well ahead of all other sectors. Both these decreases are offset by the rise in retention of workers previously employed in the bicycle manufacturing trade.

For the period as a whole, the key takeaways are that: (1) the bicycle manufacturing trade consistently recruited around 50 per cent of its labour from individuals who had not been old enough to have an occupation in the previous census; (2) retention of existing labour met a growing proportion of the labour supply, but at no stage over our period of study did this exceed a quarter of the labour force; and (3) recruitment from other Manufacturing occupations provided a notable inlet of labour, but this was always smaller than individuals 'new' (at least, since the previous census) to the workforce. This is not to say that all bicycle makers were new and with no experience of metalworking (or other relevant skills): a limited but important proportion came into the industry from related occupations, or remained as bicycle makers long term. This was likely to have imbued bicycle firms with some form of institutional knowledge, and may have facilitated their striking ability to absorb new labour.

When we turn to the composite metalworking group, one notable difference is the greater continuity of labour between censuses. This was also a fast-growing sector in the late nineteenth century, but whereas cycle making remained small-scale in 1881, the broader metalworking

sector was already well established. A sizable proportion (over a third) of composite metalworkers identified in 1891 had already been metalworkers in 1881 (Figure 4.6a), and this grew slightly over the following decades (Figure 4.6b and Figure 4.6c). This is a higher level of continuity than we find across any decennial period for the cycle industry, which probably reflects the broader nature of the composite metalworker category. As with bicycle manufacturing, the recruitment of 'new' labour was significant, although at a comparatively lower rate, starting at 40 per cent in 1891 and falling in a linear fashion to around 35 per cent by 1911. What this comparison indicates is that although both bicycle making (as a specific occupation) and metalworking (as a broader general grouping) saw expansions over the period of study, the former was particularly heavily reliant on labour new to the workforce, which had not been employed in the previous census. Rather than relying on people with existing skills, the bicycle industry equipped most of its workers with the necessary skills to produce bicycles.

At the level of the city, we find similar broad trends, but with important differences in local emphases which highlight the importance of scale. The pattern of origins for 1891 bicycle makers in 1881 is similar for Coventry and Birmingham (Figure 4.7a, Figure 4.7b), with each broadly following the national trend; however, in 1901, Coventry had begun to retain proportionally more bicycle makers than Birmingham (Figure 4.7c, Figure 4.7d). By 1911 local experience had come to play a more significant role: around 40 per cent of bicycle makers in Coventry had been in the trade in 1901 (Figure 4.7e), roughly double the figure for Birmingham (Figure 4.7f). Conversely, Birmingham saw comparably heavier recruitment of new labour, and of workers from the broader Manufacturing sector.

One explanation for this is the relative importance of the bicycle manufacturing trade in each of these places. Coventry's industry was so heavily focused on the bicycle trade that the number of bicycle makers was double that of the composite metalworker group in 1891, despite the much greater size of the latter category nationally. By contrast, bicycle makers were always in the minority in Birmingham: the Birmingham bicycle trade may have employed more people than Coventry after the turn of the century, but the industry never accounted for more than a tenth of those employed as metalworkers in other trades. Given that Coventry had such a heavy concentration of bicycle makers (relative to other similar occupations), the comparatively high retention is not unexpected.

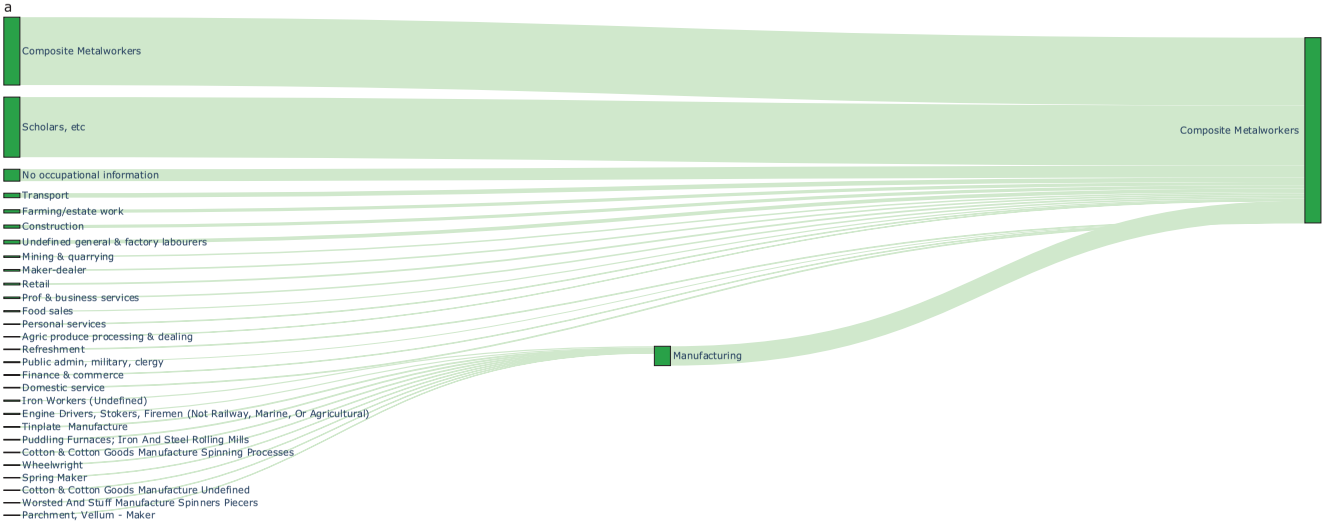


Figure 4.6a: Composite metalworkers linked to the previous census: a) 1891 to 1881, b) 1901 to 1891, c) 1911 to 1901. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

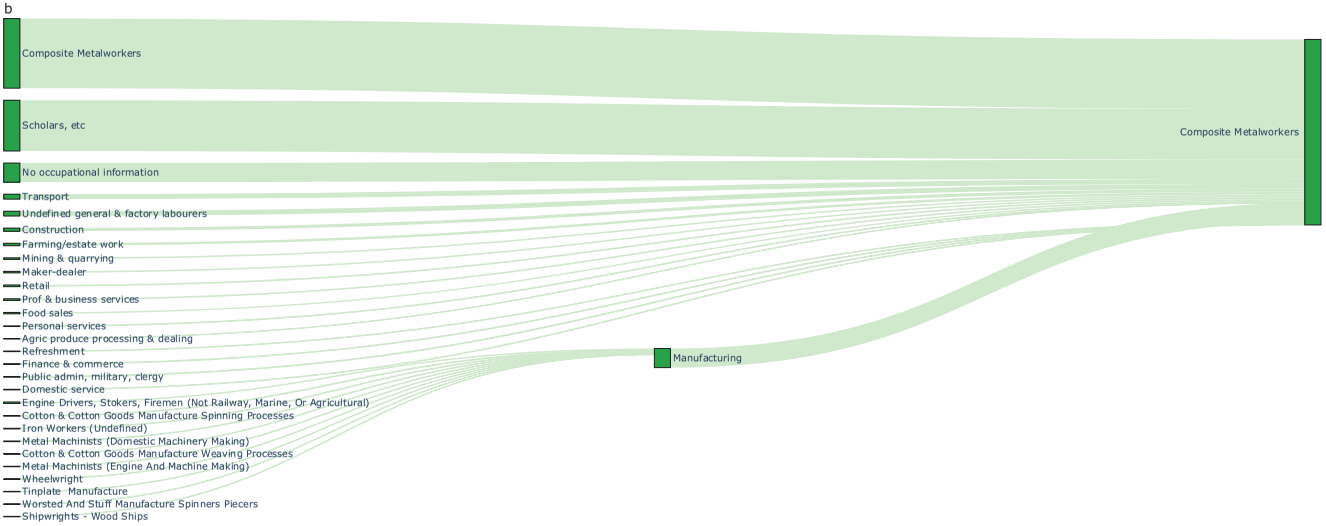


Figure 4.6b: (cont.).

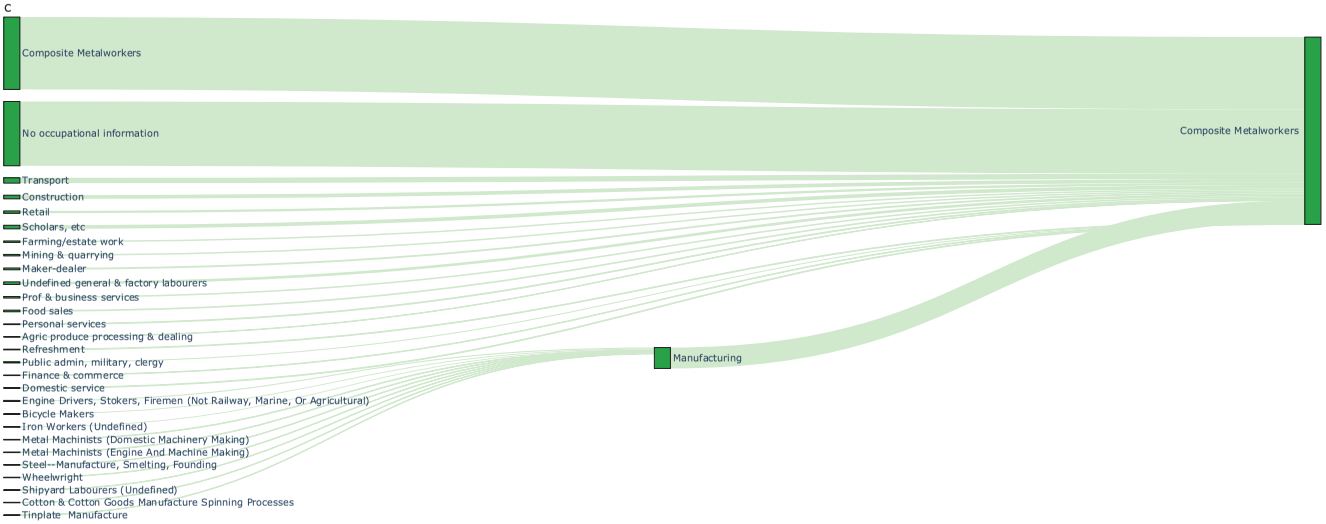


Figure 4.6c: (cont.).

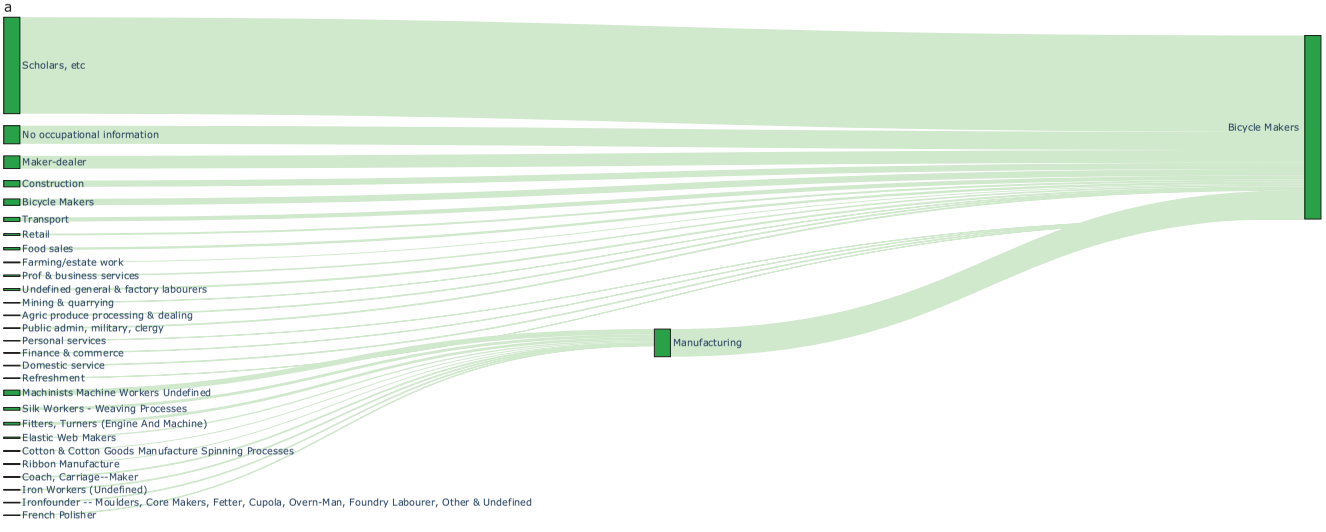


Figure 4.7a: Bicycle makers linked to the previous census: a) 1891 to 1881, Coventry; b) 1891 to 1881, Birmingham; c) 1901 to 1891, Coventry; d) 1901 to 1891, Birmingham; e) 1911 to 1901, Coventry; f) 1911 to 1901, Birmingham. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

Note: Original spellings of overn-man (oven-man) and fetter (fettler) are retained.

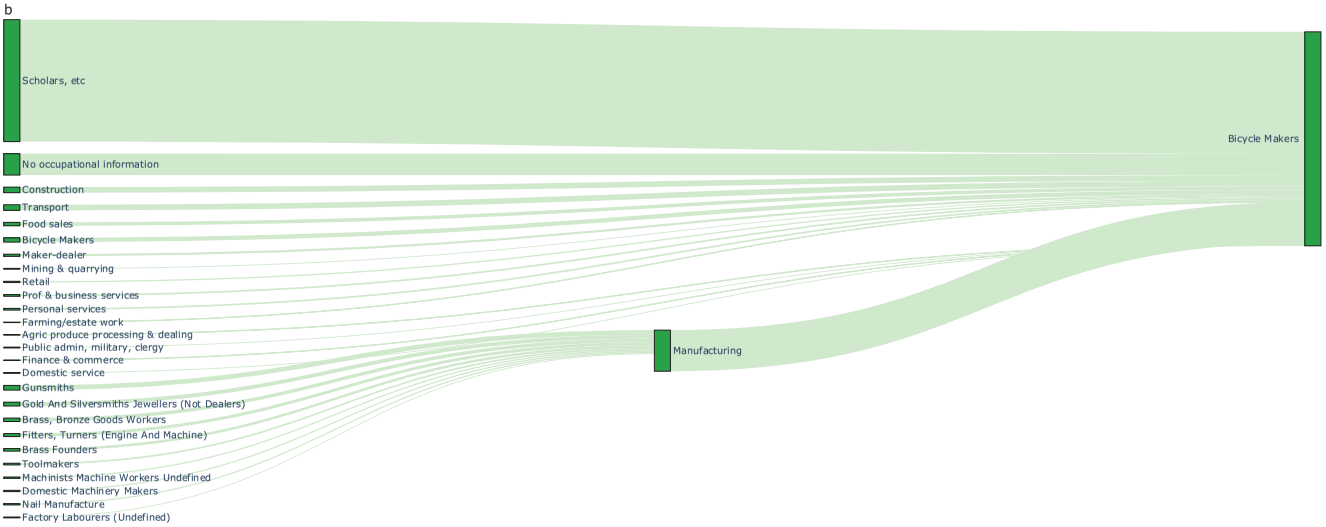


Figure 4.7b: (cont.).

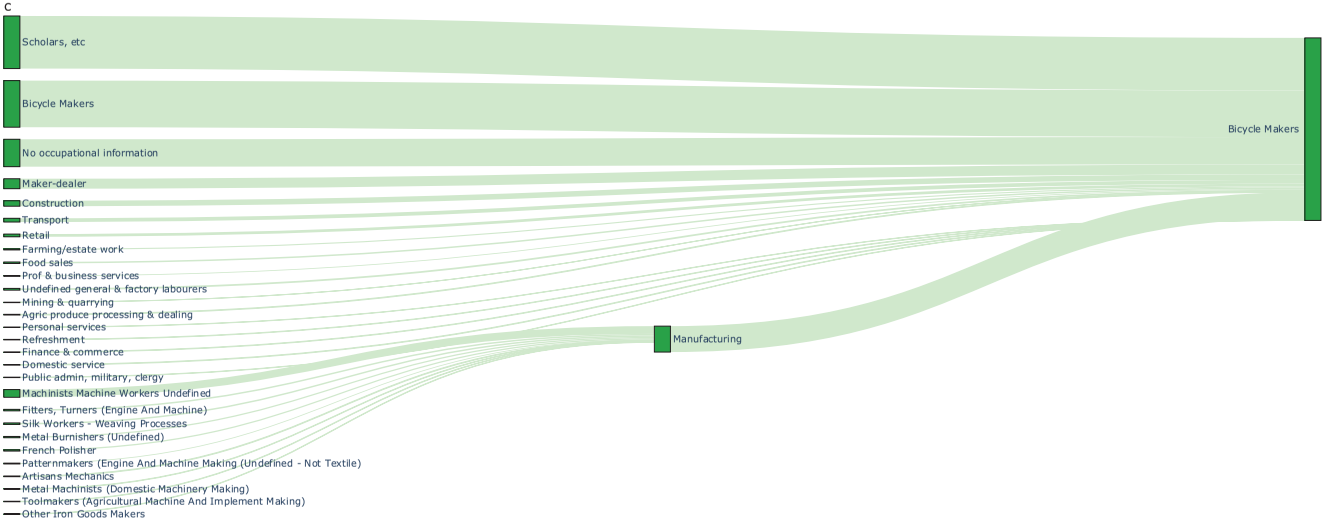


Figure 4.7c: (cont.).

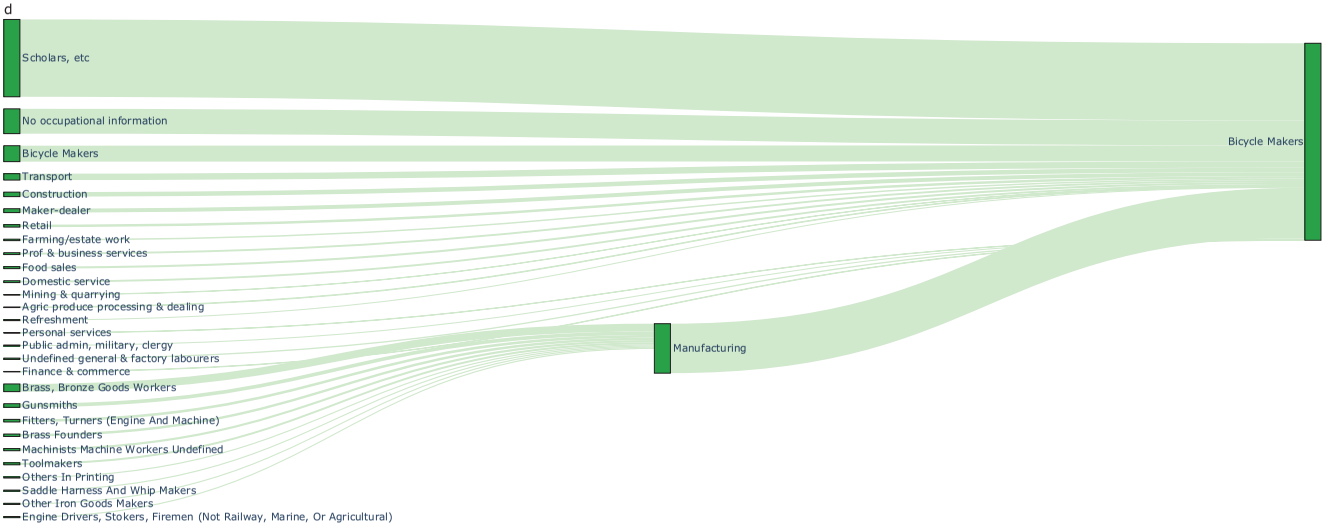


Figure 4.7d: (cont.).

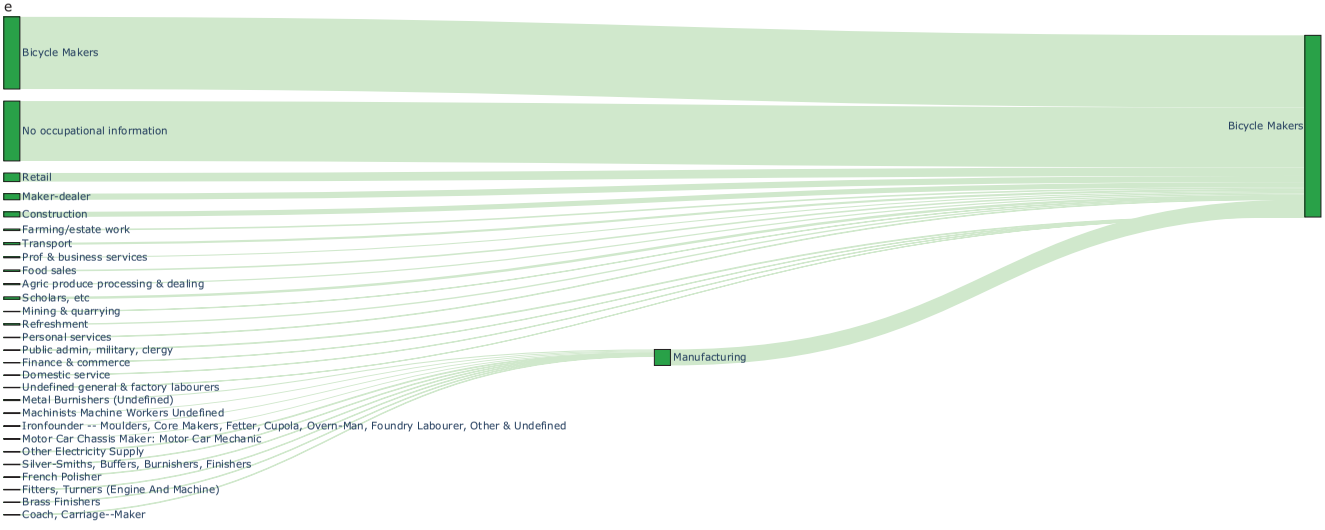


Figure 4.7e: (cont.).
Note: Original spellings of overn-man (oven-man) and fetter (fettler) are retained.

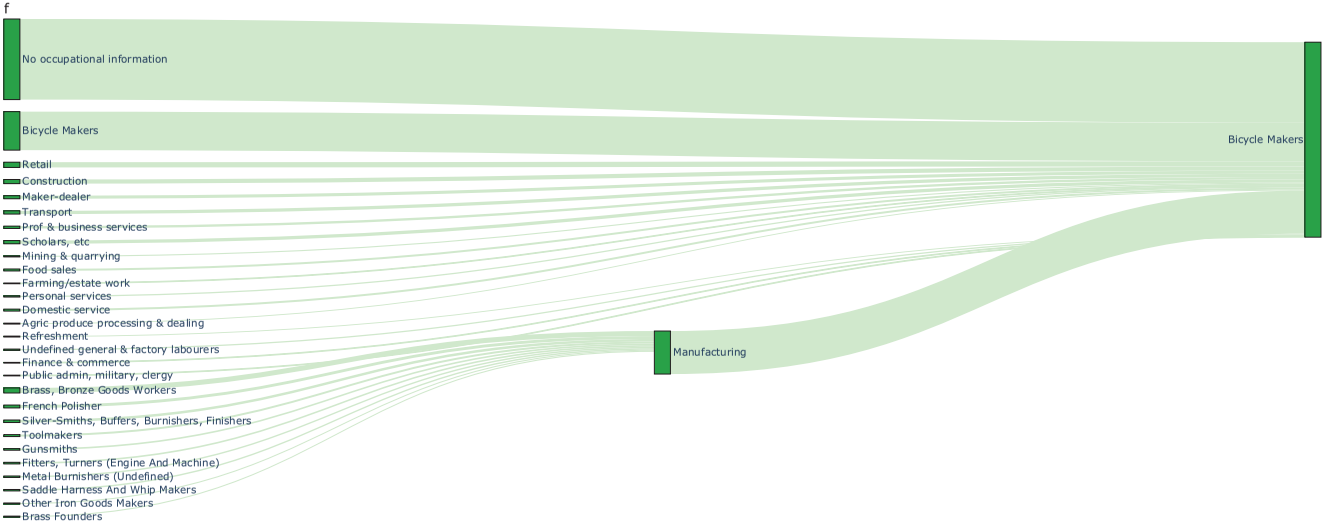


Figure 4.7f: (cont.).

Nonetheless, despite the synonymy of bicycle making with Coventry, employment in this industry was still an inherently local phenomenon. In each census year, 60 to 70 per cent of Coventry's bicycle makers had been born there (or in the adjacent parish of Foleshill), and the majority of those in the linked sample had also been living there ten years earlier. The situation was very similar for Birmingham – 60 to 70 per cent of bicycle makers had been born there, and most were living in Birmingham at the previous census. Given the relative proximity of the two cities and their status as the major bicycle manufacturing centres, it is perhaps surprising that there appears to have been relatively minimal exchanges of labour between them. Even during the boom years of the 1890s, the employment effects of the industry's rapid expansion appear to have been intensely local. In what is captured by the census, we find little evidence of large-scale in-migration of labour, skilled or otherwise.

The relationship between specialised bicycle making and more general metalworking was very different in Coventry by 1911. The former still outnumbered the latter in 1901 – although this ratio had been reduced to around 1.5 – but had heavily inverted by 1911. Indeed, the period between 1901 and 1911 had seen both a reduction in the number of bicycle makers by a little less than a half, and the tripling in the number of composite metalworkers, such that the number of those employed in bicycle manufacturing was now around a quarter of those employed in other metalworking occupations for that year.

The increase in the overall number of composite metalworkers for 1911 in Coventry was not just driven by one industry. Occupations which formed part of (or were related to) motor vehicle manufacturing certainly account for some of this change, but the data also indicate a notable rise in those employed as railway signal switch turntable fitters, artisan mechanics and toolmakers (among others). Given the oversaturation in bicycle manufacturing, we might also expect to see a flow of displaced bicycle makers into other branches of metalworking in Coventry between 1901 and 1911, but at around 13 per cent of the latter in 1911 having been employed as the former in 1901, this is actually no higher, proportionally, than in the previous decade (Figure 4.8). This is a theme to which we return more fully in the 'Occupational Destinations' section.

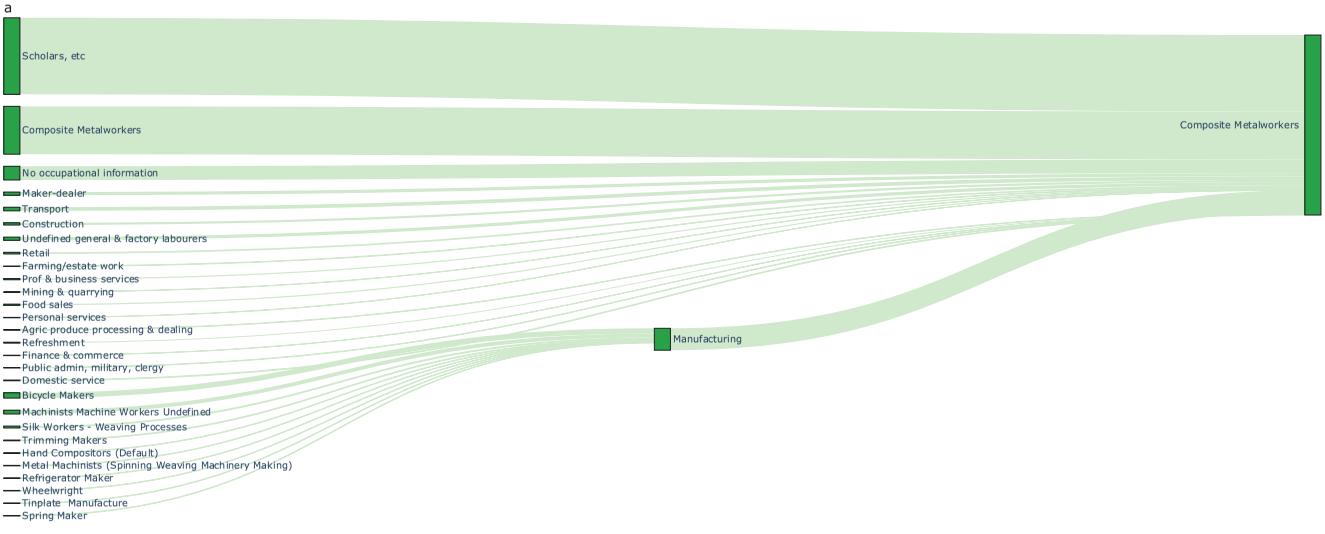


Figure 4.8a: Composite metalworkers linked to the previous census: a) 1891 to 1881, Coventry; b) 1891 to 1881, Birmingham; c) 1901 to 1891, Coventry; d) 1901 to 1891, Birmingham; e) 1911 to 1901, Coventry; f) 1911 to 1901, Birmingham. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

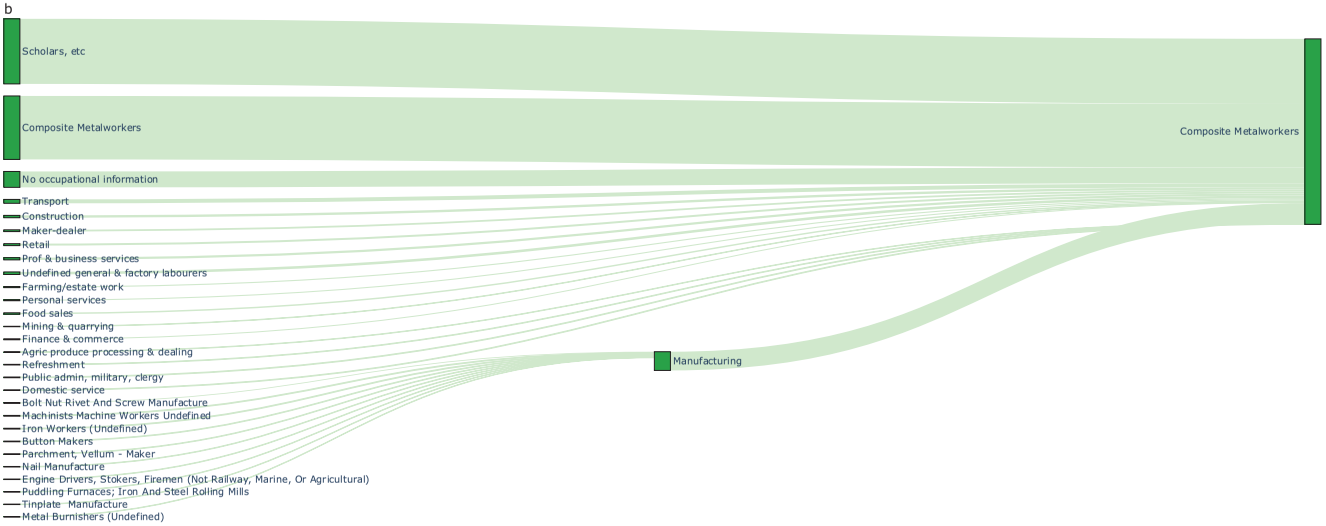


Figure 4.8b: (cont.).

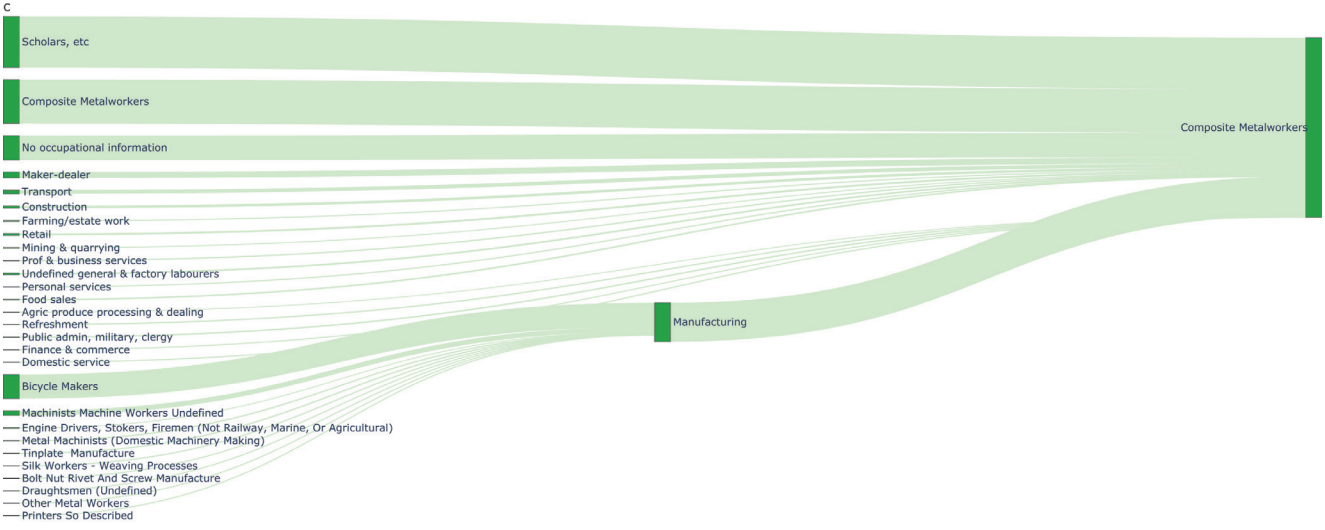


Figure 4.8c: (cont.).

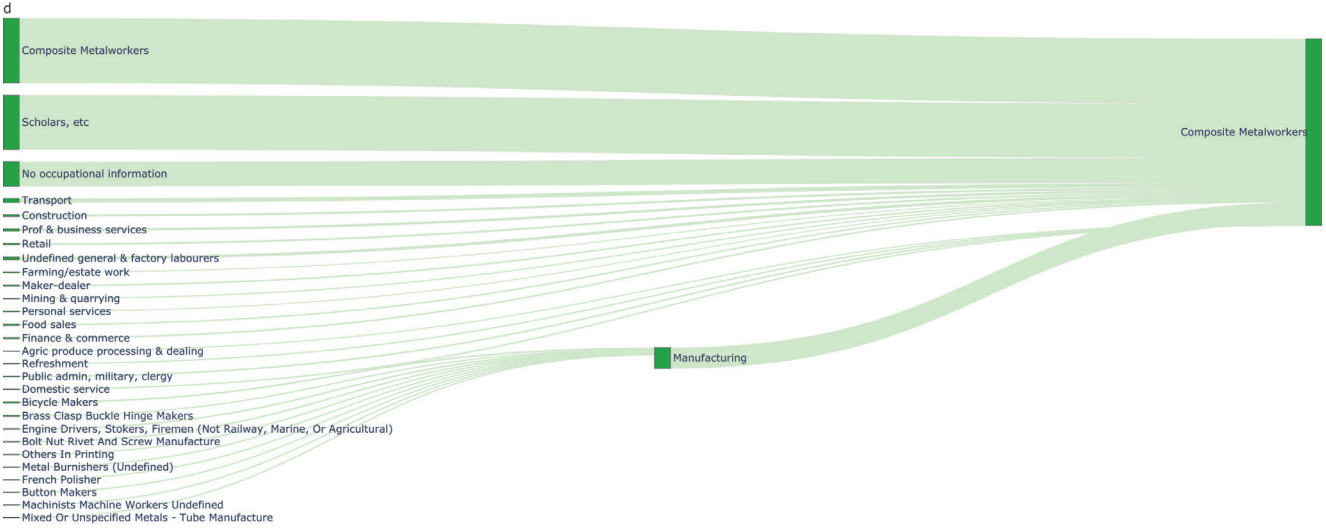


Figure 4.8d: (cont.).

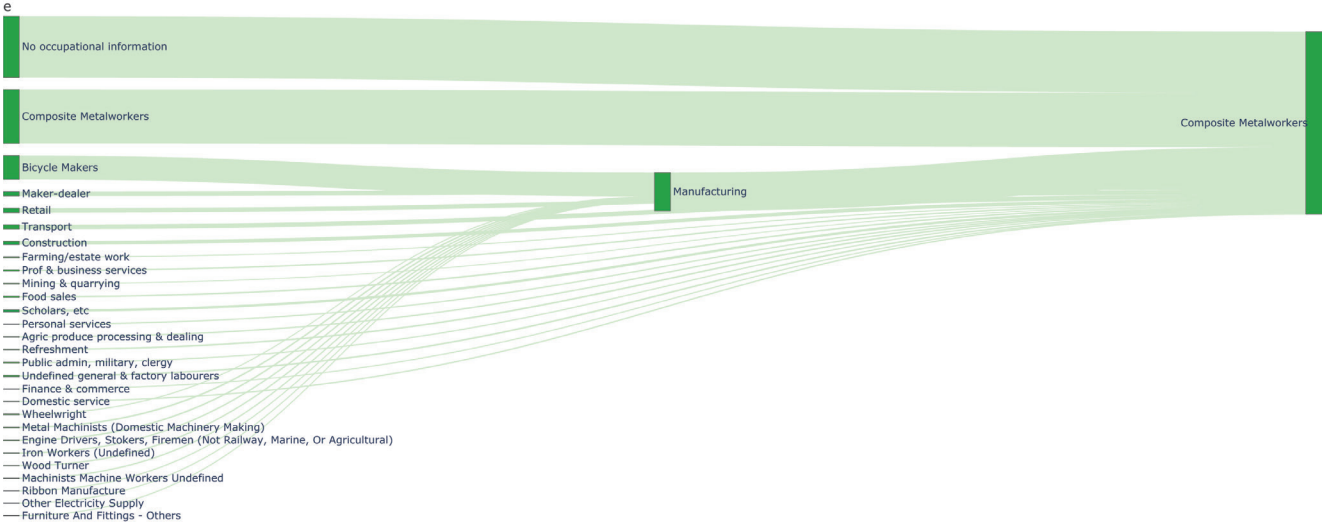


Figure 4.8e: (cont.).

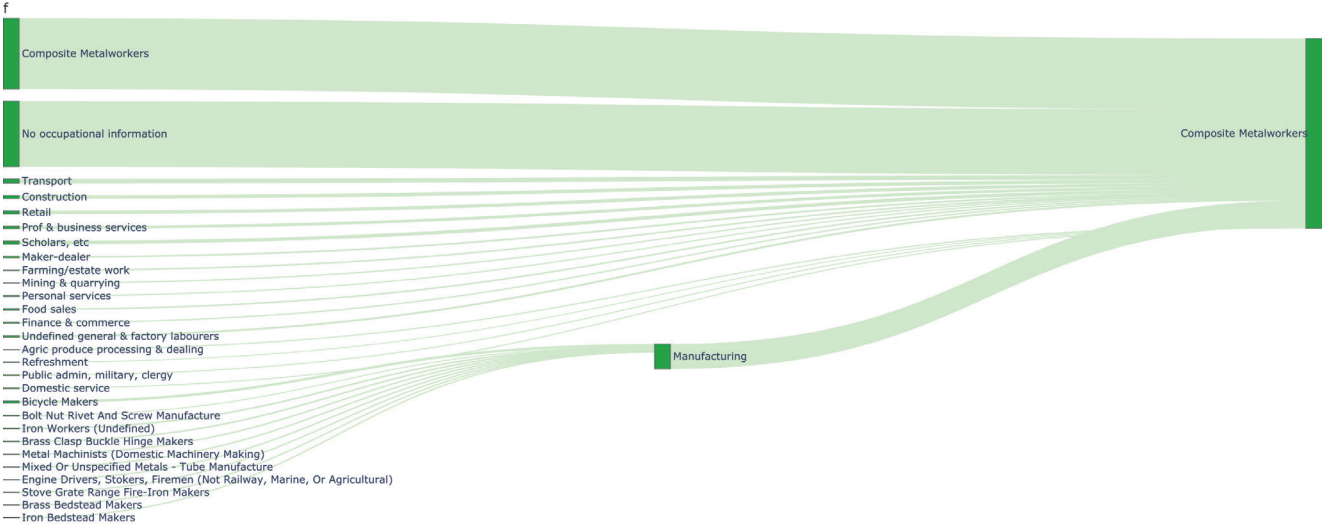


Figure 4.8f: (cont.).

Analysis: occupational origins redux (where did their ancestors *come from?)*

If, as we find here, there was no great transition of workers from occupations associated with proto-bicycle manufacturing industries combined with only limited in-migration, this raises the question of how we should understand the radical sectoral shifts in employment seen in places – like Coventry – that were at the heart of the national cycle boom. Given the cycle industry's heavy reliance on 'new' labour, a useful way to understand this sectoral shift is to explore the type of households from which bicycle makers were drawn. It is here that the *nominal linkage* method facilitates a particularly insightful new approach. Observing new labour coming from households with some previous connection to gun-making or sewing machine manufacturing would support the idea of an intergenerational skills transfer or occupational legacy. On the other hand, if this new labour came from households which were seemingly unrelated to the technological convergence that fostered the bicycle trade – such as those experiencing economic downturns or other forms of industrial displacement – then this would lend credibility to the idea that skills for bicycle making were learned largely on the job.

We can consider household characteristics using the same linked sample as discussed above. By linking those identified as 'new' labour back to the previous census and then considering the recorded occupation of the household head, we are able to establish the occupational role in which the latter person worked. We select 'household head' over 'father' as not all people had fathers living in the household, and we assume that the nature of the household was best represented by the person recorded as its nominal head. We can also extract the relationship that each individual identified as a new bicycle maker in the following census had to the household head. As might be expected, most of these were classified as 'Son', although a variety of other relationships were recorded.

As in the previous section, we break down the household head's occupation on two levels: the sector and the specific occupation.[42] It is important to include both because, as we see from Figure 4.9a, although the most frequent sector of employment for the household head was Manufacturing, the most frequent occupation was 'Watch and Clock Maker' (which belonged to the Maker-Dealer sector). This was true for the census-pairs of both 1881/1891 and 1891/1901, but not 1901/1911. By then, 'Watch and

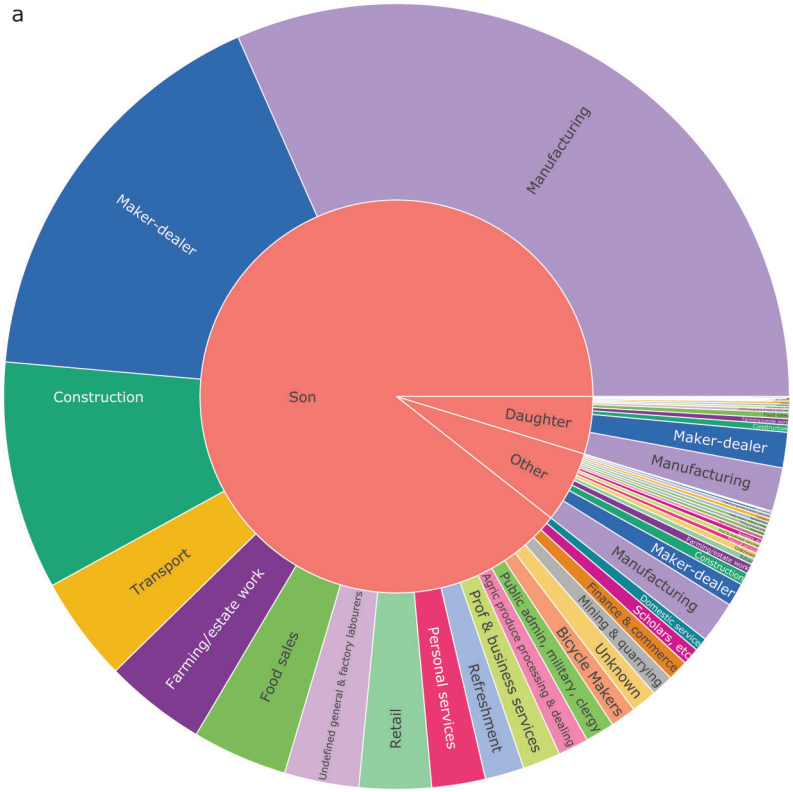


Figure 4.9a: The occupation of household heads, where a member of the household had no occupation and was a bicycle maker in the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service. An expanded version of this visualisation with an additional ring for occupations is available in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

Clock Makers' family backgrounds had sharply declined, and the most frequent recorded occupation for heads of household was bicycle maker.

In the early years of the bicycle manufacturing industry, the most prominent Manufacturing occupation for workers' heads of household was 'Silk Workers – Weaving Processes' followed by 'Elastic Web Makers'. However, there were also other notable occupations from other sectors, such as 'Carpenter, Joiner' (Construction) and 'General Labourers' (Undefined General and Factory Labourers). As the period progressed, the prominent occupations within Manufacturing altered somewhat, with

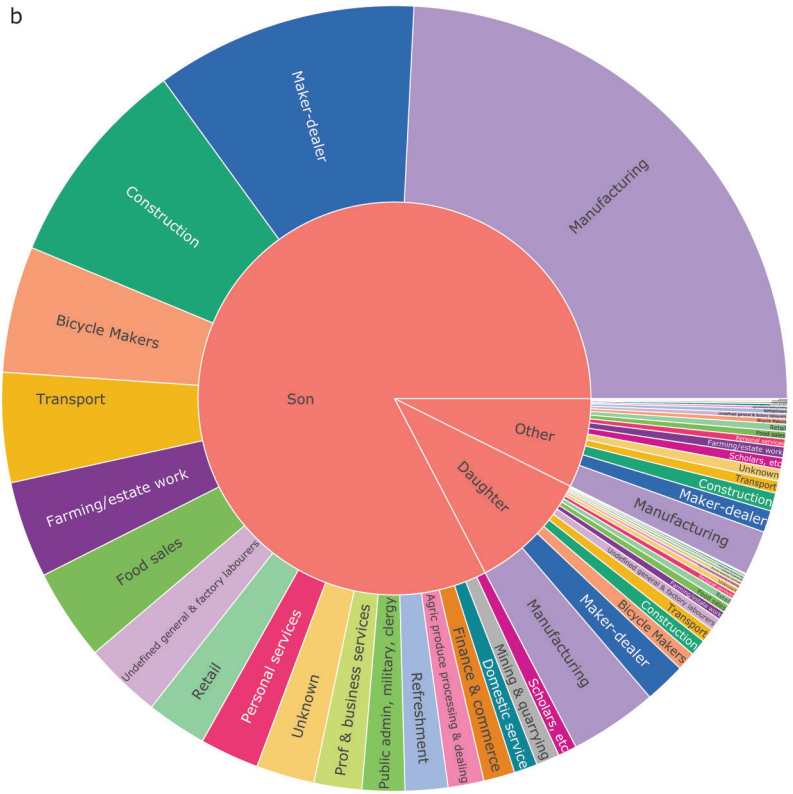


Figure 4.9b: (cont.).

much more parity among the specific occupations featured (Figure 4.9b and Figure 4.9c).

This is not to say that occupations associated with the history of bicycle manufacturing firms did not feature – 'Gunsmiths' were fairly prominent among those who came from Manufacturing households throughout the period – but they did not do so in particularly great numbers. Rather, the commonalities in the leading occupations seem to have been industries experiencing their own economic downturns coincidental to the rise of the bicycle trade. In Coventry, silk weaving and, to a lesser extent, watch and clock making, previously major local industries, were hit hard by overseas competition from the 1860s and many children of workers in these trades consequently looked elsewhere for employment.[43]

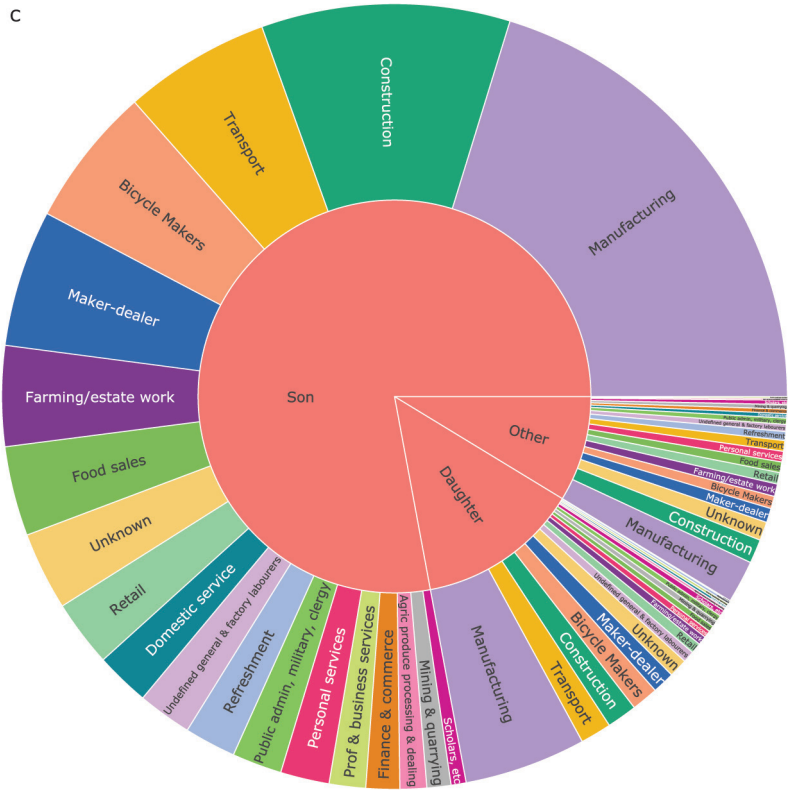


Figure 4.9c: (cont.).

Indeed, the major trends in the national picture for 1881/1891 (Figure 4.9a) appear to be driven by the situation in Coventry. As shown by Figure 4.10a, the prominence of 'Watch and Clock Makers' and 'Silk Weavers – Weaving Processes' is exacerbated at the city level – so much so that while the Manufacturing sector remains the single largest contributor, the margin between this and Maker-Dealers is all but gone. And while the importance of 'Watch and Clock Makers' is lower at the national level thereafter, it remained prominent within the households of those entering the trade in Coventry in 1891/1901 (Figure 4.10b) and – albeit less so – in 1901/1911 (Figure 4.10c). Much as at the national level, bicycle makers themselves also became increasingly important – but again, the scale of this importance is exaggerated in the case of Coventry. The concentration of the industry in Coventry

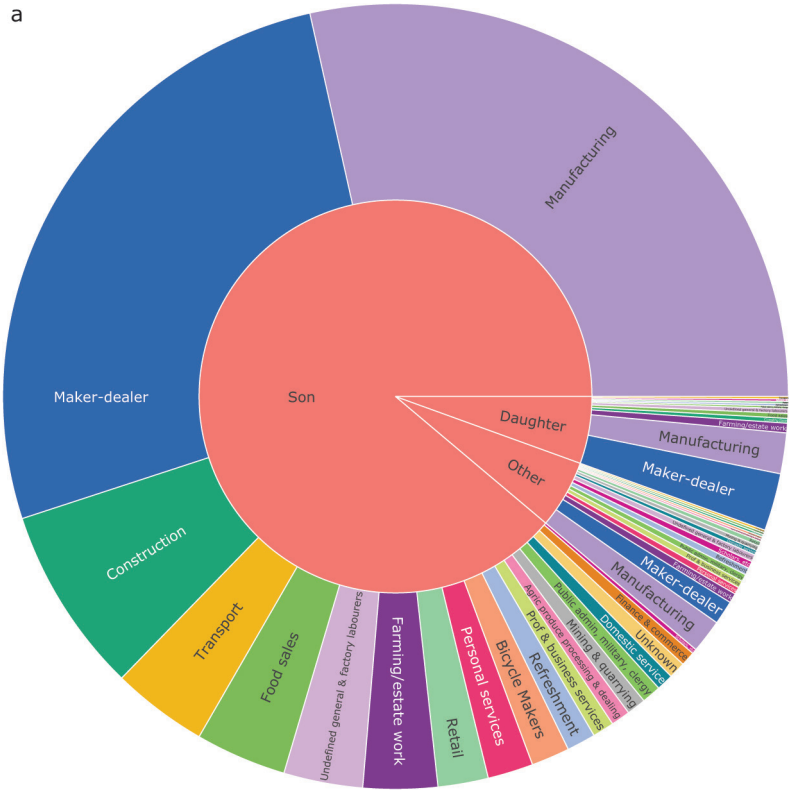


Figure 4.10a: The occupation of household heads, where a member of the household had no occupation and was a bicycle maker living in Coventry at the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service. An expanded version of this visualisation with an additional ring for occupations is available in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

therefore meant that it exerted a sizable influence over the national picture, and thus the inflows of labour from industries experiencing economic downturns in Coventry played a significant role in the trade as a whole.

Just as for occupational flows, at the national level, the longer established nature of the broader metalworking sector also shows through in household heritage. In 1881/1891, metalworkers were already the most frequent household head from which new metalworking labour was drawn (Figure 4.11a). Manufacturing (and to a lesser extent Construction and Transport, as well as 'General Labourers') still played important roles,

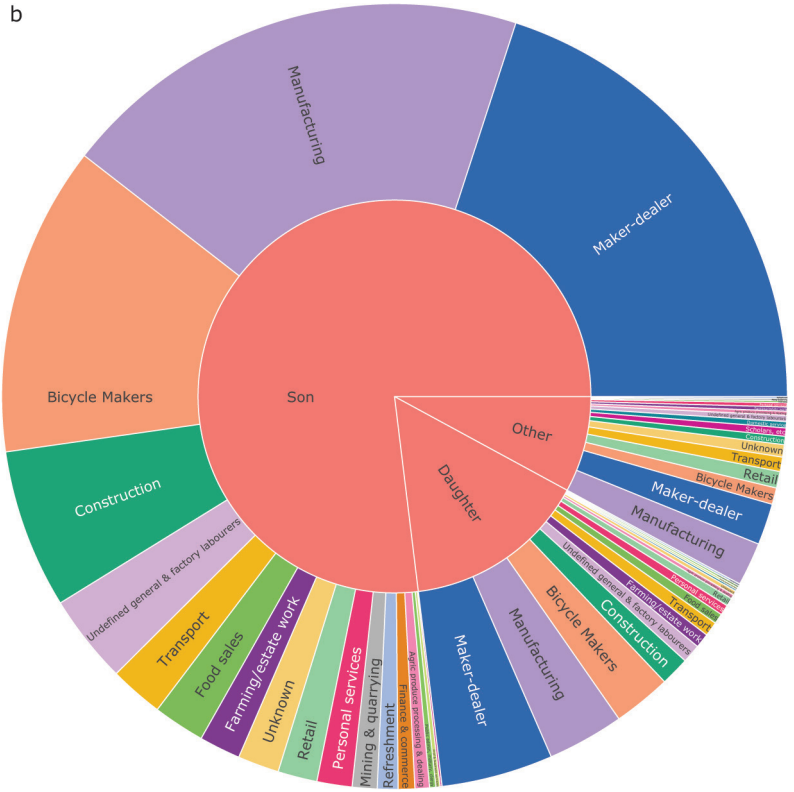


Figure 4.10b: (cont.).

but recruitment from the households of Maker-Dealers is notable by its absence. The picture was largely the same in 1891/1901 (Figure 4.11b) and 1901/1911 (Figure 4.11c), although the increasing prominence of 'Daughters' is notable in the latter year – who, relative to 'Sons', were slightly more heavily drawn from Manufacturing households not included in the composite metalworkers grouping (something we also see replicated in bicycle makers).

However, this story is once again influenced by scale. The relative preponderance of composite metalworkers outside of Coventry means that the national picture is much less heavily influenced by its household heritage than that of bicycle makers. In fact, comparing Coventry's composite metalworkers to its bicycle makers yields a remarkably similar picture

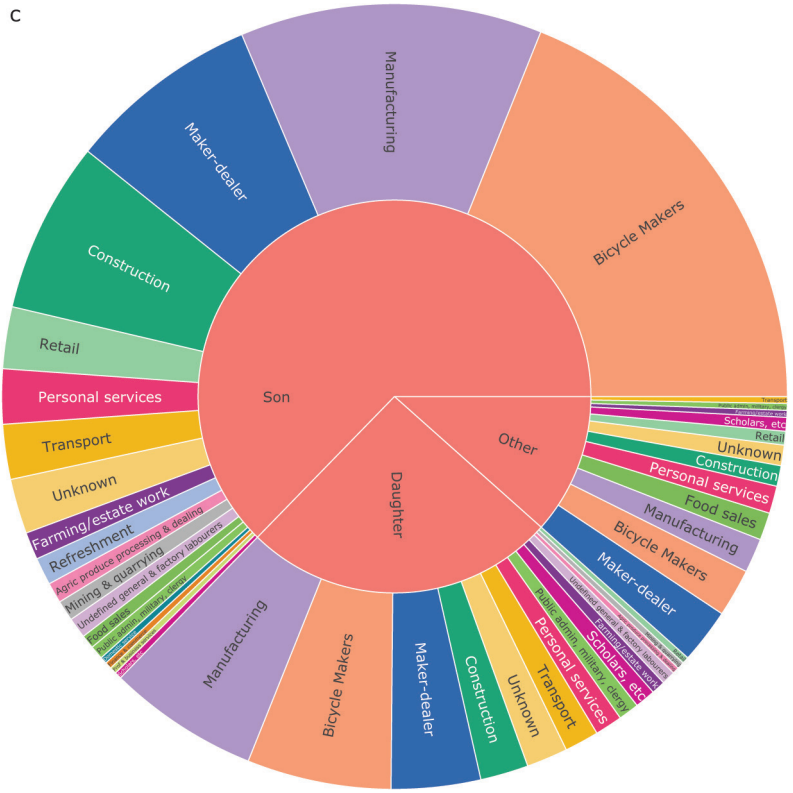


Figure 4.10c: (cont.).

(Figure 4.12). The longevity of metalworking occupations still exists in 1881/1891 in ways not seen for bicycle making, but approximately half of new labour in both cases was drawn from other Manufacturing or Maker-Dealers with 'Watch and Clock Makers' and 'Silk Workers – Weaving Processes' again the most prominent occupations. A similar story is presented in 1891/1901 – although the sons of bicycle makers were by then the most common household heritage grouping recruited into the composite metalworking group of occupations.

As previously established, the period between 1901 and 1911 was one of change in the bicycle manufacturing/metalworking occupational structure in Coventry. Accordingly, what is particularly interesting is that this did not lead to a radical shift in the household heritage of new

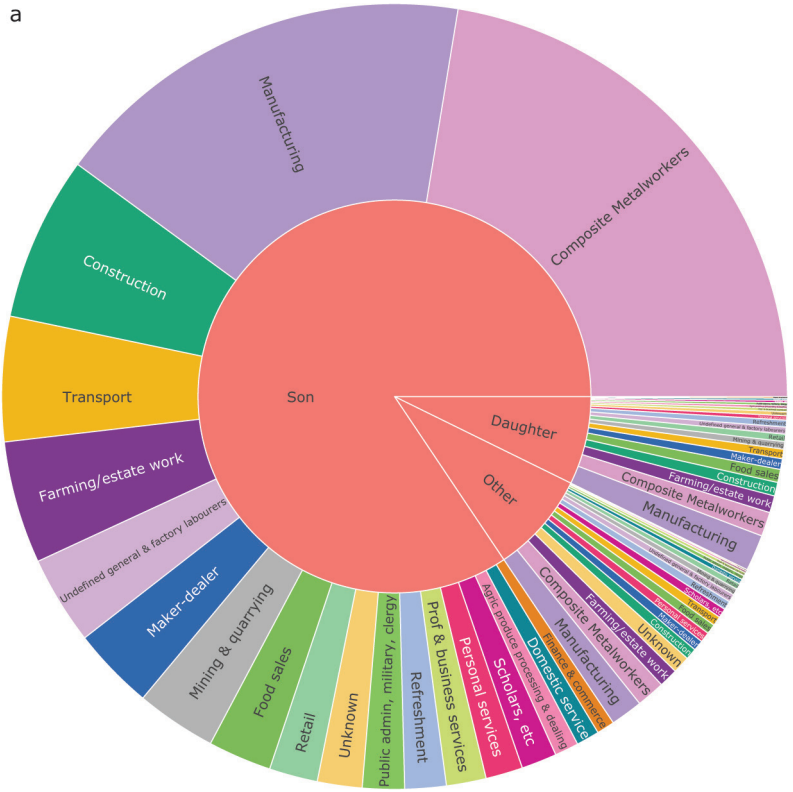


Figure 4.11a: The occupation of household heads, where a member of the household had no occupation and was a composite metalworker in the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service. An expanded version of this visualisation with an additional ring for occupations is available in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

labour. Proportions are comparable across sectors, and bicycle makers and 'Watch and Clock Makers' remain the most prominent heritage occupations (Figure 4.10c and Figure 4.12c). Again, what is noticeable is the growing number of 'Daughters' in comparison to 'Sons', which – in Coventry – was much more prominent for bicycle makers than for other metalworking trades. This appears to have been due to the combined effect of a shrinkage in the numbers of male bicycle makers and a growth in female ones – although comparing the sons and daughters groupings demonstrates similar household origins.

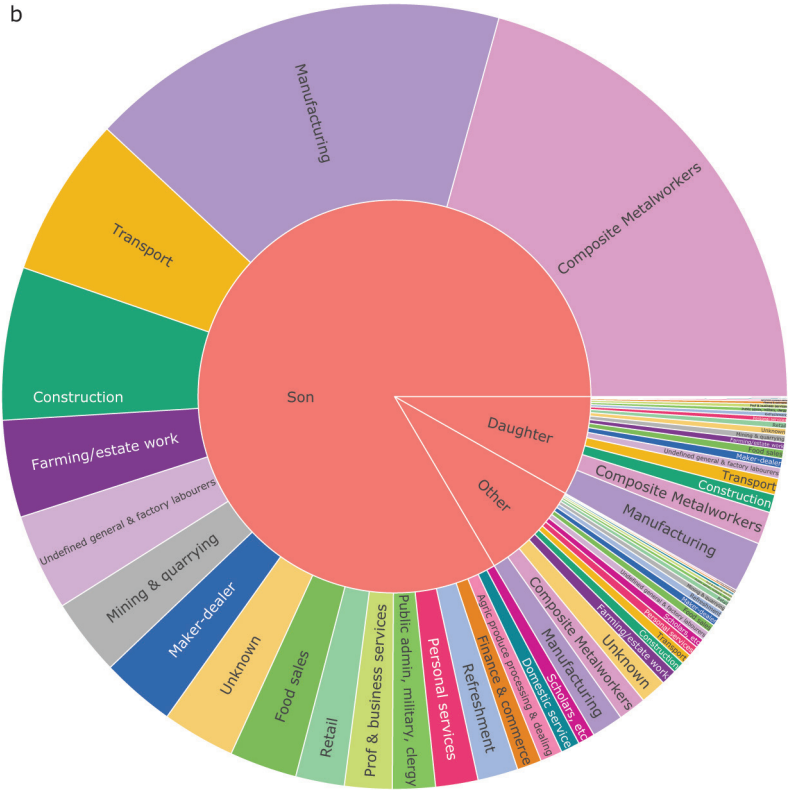


Figure 4.11b: (cont.).

In combination, these analyses yield several key takeaways. First, there appears to be some form of hereditary element in both bicycle making and broader metalworking. In the most simple terms, sons often followed their fathers into their occupation. Nonetheless, the strength of this trend should not be overstated, and recruitment from the households of occupations undergoing their own economic downturns seems to have been the over-riding pattern. On the basis of the evidence presented here, bicycle makers do not appear to have established themselves as an occupational group distinct from metalworkers in general – with considerable overlap in terms of the households from which each recruited new labour. However, this trend can be disguised when only considering national-level data:

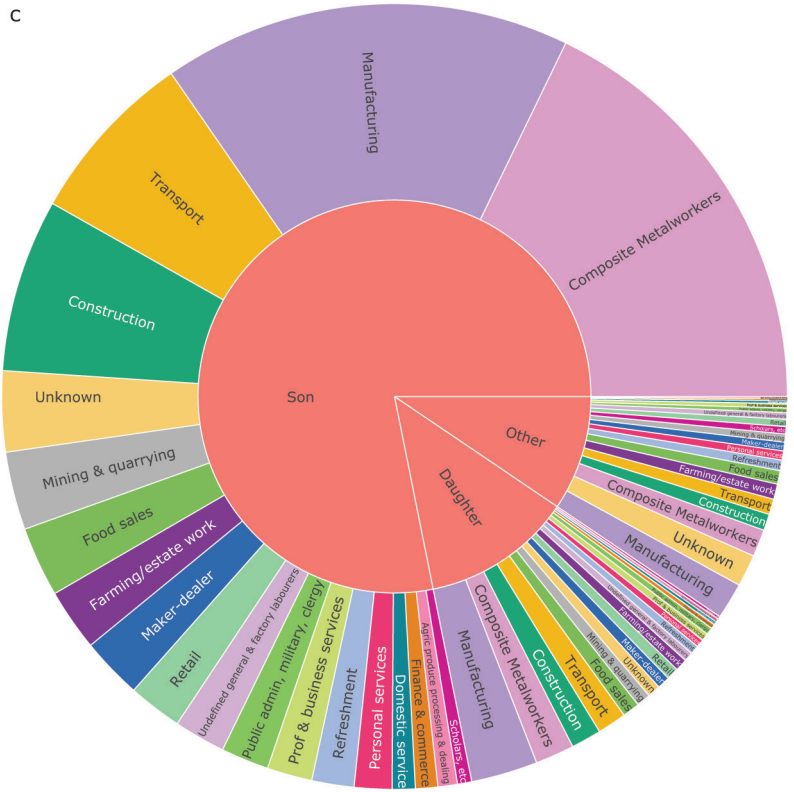


Figure 4.11c: (cont.).

the significant clustering of bicycle manufacturers requires a local-level comparison. In short, the two industries displayed similar (but distinctive) recruitment patterns in Coventry, but this city was so important to national employment data for bicycle makers that this creates the appearance of more pronounced sectoral differences nationally.

Analysis: occupational destinations (where did they go?)

The previous sections highlighted the role of labour new to the workforce in both the specific occupation of bicycle making, and composite metalworking grouping more generally. However, the entry point for young workers was not necessarily their permanent destination. Therefore,

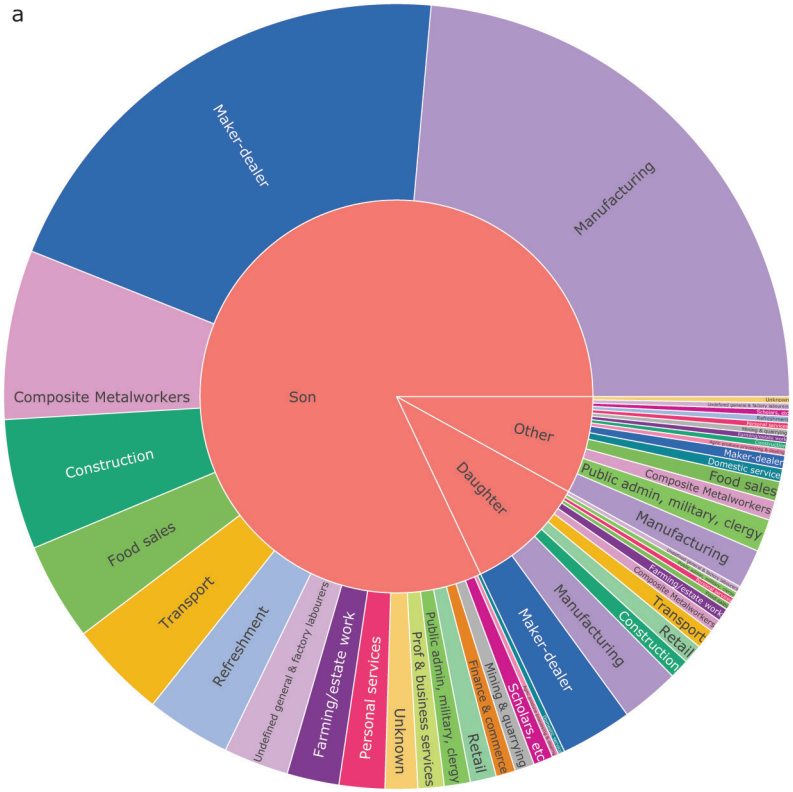


Figure 4.12a: The occupation of household heads, where a member of the household had no occupation and was a composite metalworker living in Coventry at the following census: a) 1881/1891, b) 1891/1901, c) 1901/1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service. An expanded version of this visualisation with an additional ring for occupations is available in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

in this section we reverse the chronology, and instead of evaluating the origins of labour (by linking individuals to the previous census), we investigate its destinations (and consider what people were doing ten years after they were identified as bicycle makers). If – as it appears – bicycle manufacturing largely relied on new labour, what did these individuals go on to do ten years later? Or, in other words, were bicycle makers only qualified to be bicycle makers (or other closely related occupations), or did they have useful skills for other occupations as well? One associated question relates to the influence of the motor industry.

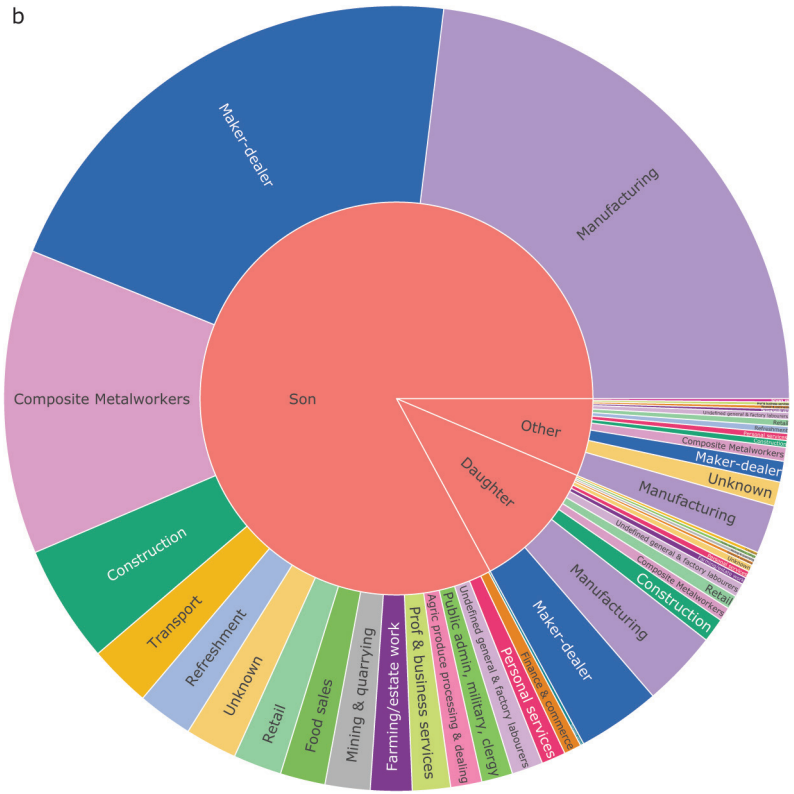


Figure 4.12b: (cont.).

As noted, many Midlands firms pivoted from bicycle manufacturing to the production of motor vehicles, but – much as the previous progression from gun-making and sewing machine manufacturing was not necessarily reflected in those subsequently employed as cycle workers – it is not clear if individuals employed as bicycle makers accompanied their employers through this transition.

At a national level, we again see that bicycle makers demonstrate a related but distinct pattern from metalworkers more broadly. Of the bicycle manufacturers identified in 1891, around 40 per cent were still working in the trade ten years later (Figure 4.13a). A further 30 per cent had moved into the broader Manufacturing sector, most commonly as fitters/turners or toolmakers, but there was not a particularly dominant occupation.

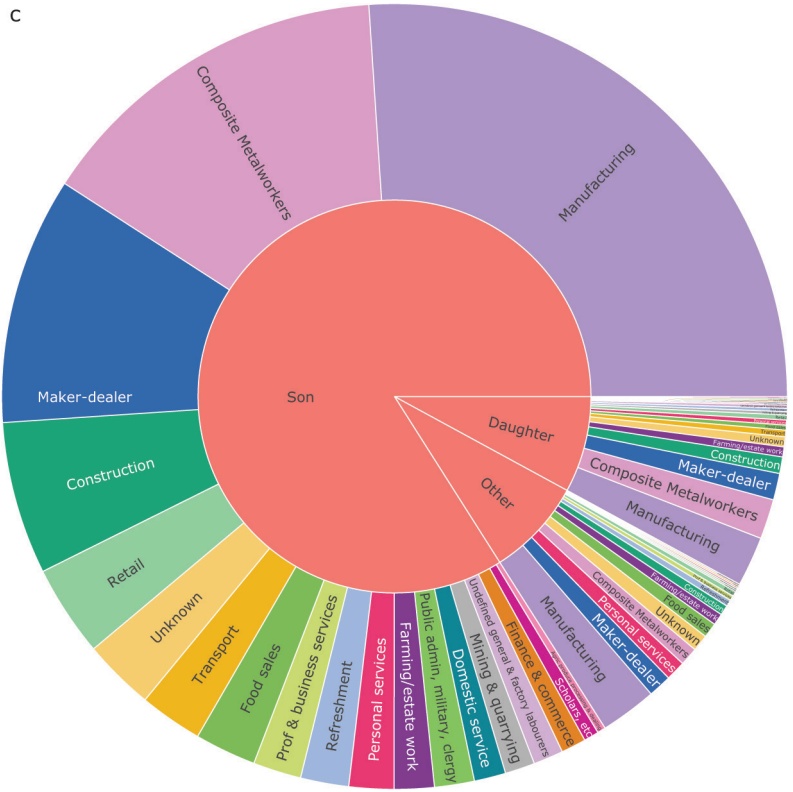


Figure 4.12c: (cont.).

Of the other sectors, Construction and Retail saw the most notable flows, but neither exceeded ten per cent of the total individuals employed.

The changes in the industry were far more apparent between 1901 and 1911 (Figure 4.13b). Only around a quarter of bicycle makers in 1901 were still working as such in 1911, and almost 40 per cent had moved on to other Manufacturing occupations. Within this, the switch to motor vehicle manufacturing becomes increasingly clear – a quarter of those leaving for Manufacturing left for motor car making, and in practice this percentage was probably higher, once jobs relating to motor vehicle making but not definitively labelled as such are considered. Nonetheless, it is important not to overstate the transition from bicycle to motor vehicle making in these years. While the flow from the former to the latter is certainly present, it only accounted for around ten per cent of workers overall.

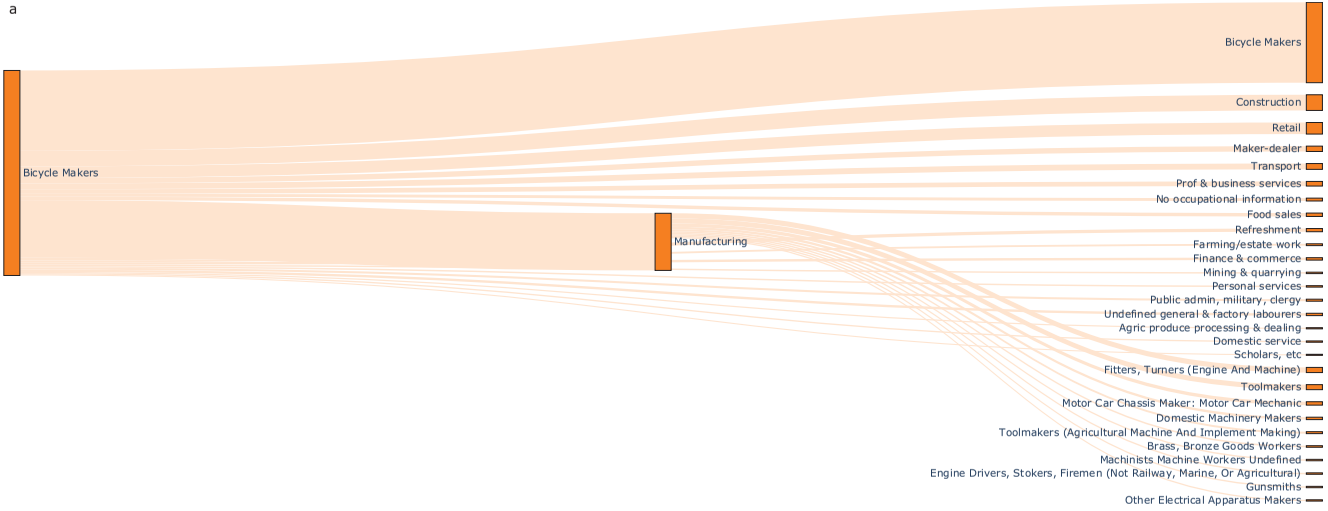


Figure 4.13a: Bicycle makers linked to the following census: a) 1891 to 1901, b) 1901 to 1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

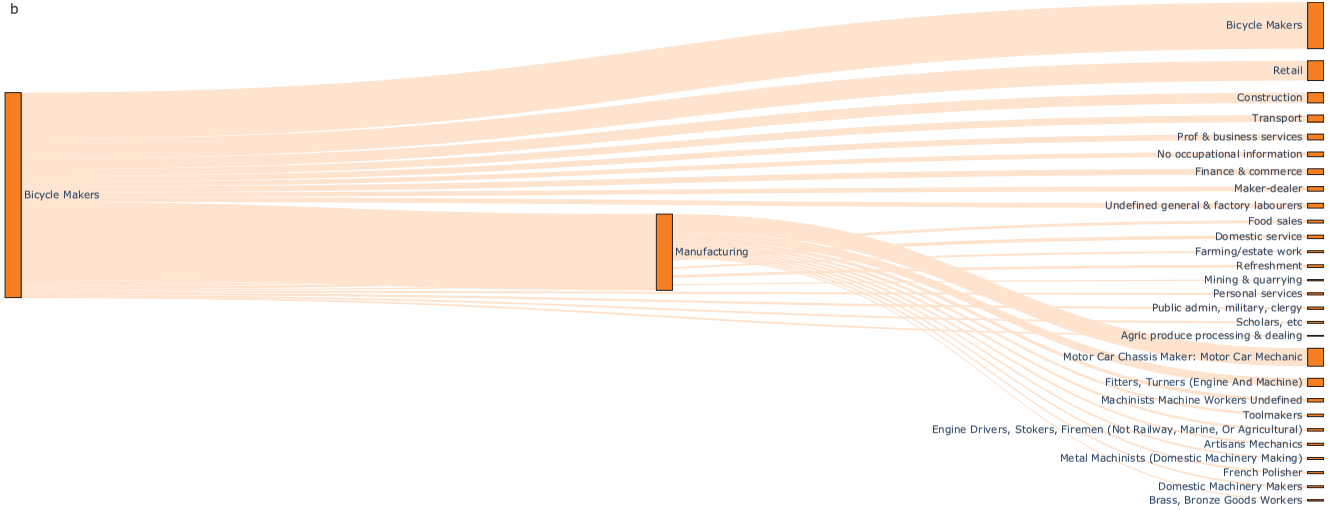


Figure 4.13b: (Cont.).

This is likely to have progressed in the years after 1911, but it is not yet possible to say this with certainty without the inclusion of the 1921 census data.

For composite metalworkers, the picture is much more static. Of these metalworkers in 1891, 60 per cent were still metalworkers in 1901, and a further 15 per cent had left for other Manufacturing occupations (Figure 4.14a). The remainder show relatively little by way of coherent pattern, and were split across various sectors in fairly small numbers. This was much the same ten years later (Figure 4.14b): around 55 per cent of composite metalworkers in 1901 remained composite metalworkers in 1911 (despite the growth in their overall numbers), and around 15 per cent had moved to other Manufacturing employment.

The comparatively high level of retention in composite metalworking suggests that those who worked in metalworking jobs tended to remain within that general field (rather than, for example, moving between other Manufacturing occupations which might have been broadly equivalent in terms of status) and therefore had likely developed a somewhat specialised skillset. The relatively lower retention for a specific occupation like bicycle making suggests that this skillset was not so specialised as to restrict these workers to a very narrow occupational focus – but was sufficient to create a certain degree of occupational stickiness.

At a finer spatial scale, the key question is how bicycle makers in Coventry reacted to shifts in the industry. The answer is that this largely resembled the aggregate national level across the trade. In both Coventry (Figure 4.15a) and Birmingham (Figure 4.15b), and other places (such as Wolverhampton) between 1901 and 1911, 20 to 30 per cent of bicycle makers remained as such in the latter year, while around 40 per cent left for other manufacturing (including other metalworking trades).

In summary, we see that the majority of metalworkers continued to be recorded as metalworkers in later censuses. While the rate at which bicycle makers were recorded as continuing in the bicycle manufacturing trade was lower, this is to be expected with the greater occupational specificity and more limited employment opportunities (although the latter was counteracted by the geographical concentration of the industry, at least to some degree). Nonetheless, we still observe significant retention in those employed by the bicycle trade.

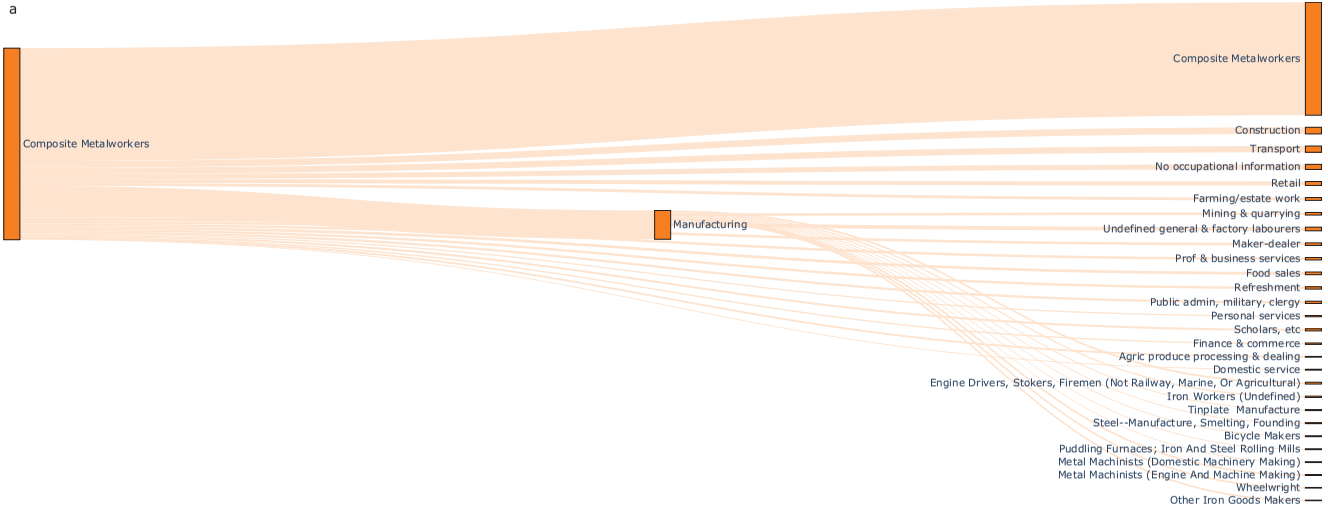


Figure 4.14a: Composite metalworkers linked to the following census: a) 1891 to 1901, b) 1901 to 1911. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

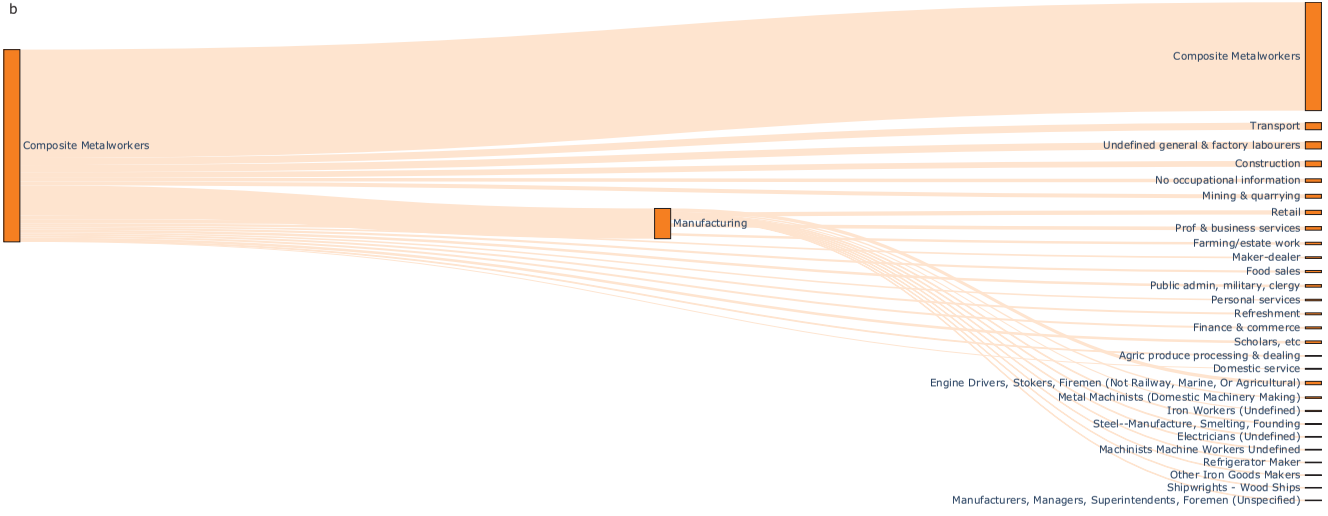


Figure 4.14b: (Cont.).

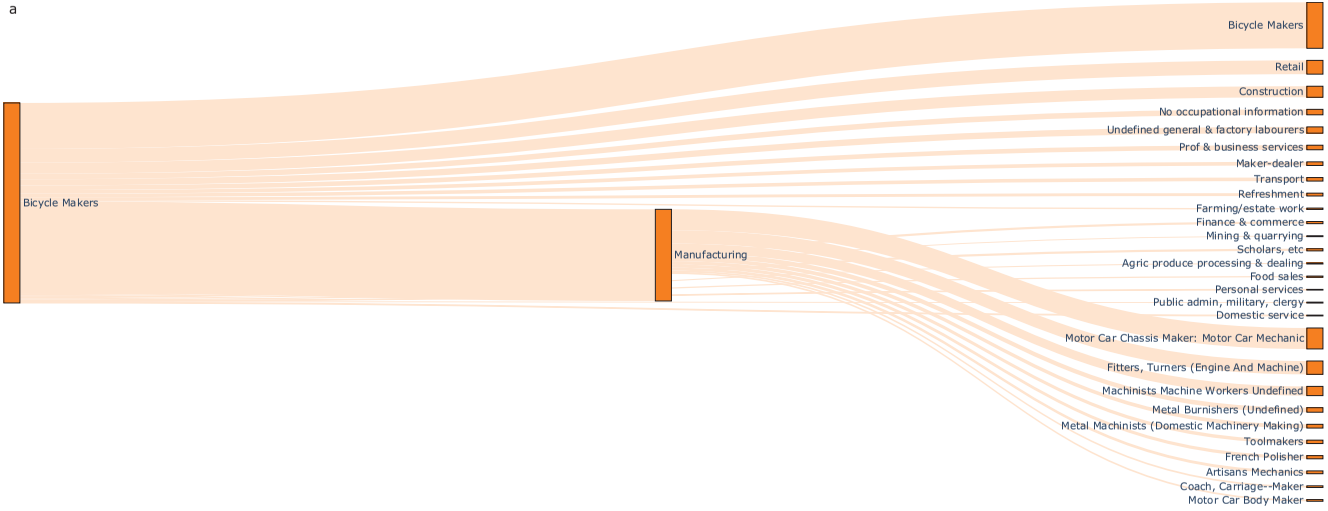


Figure 4.15a: Bicycle makers in 1901 linked to the 1911 census in: a) Coventry, b) Birmingham. Credit: Guy Solomon. Underlying data derived from K. Schurer and E. Higgs (2020). UK Data Service.

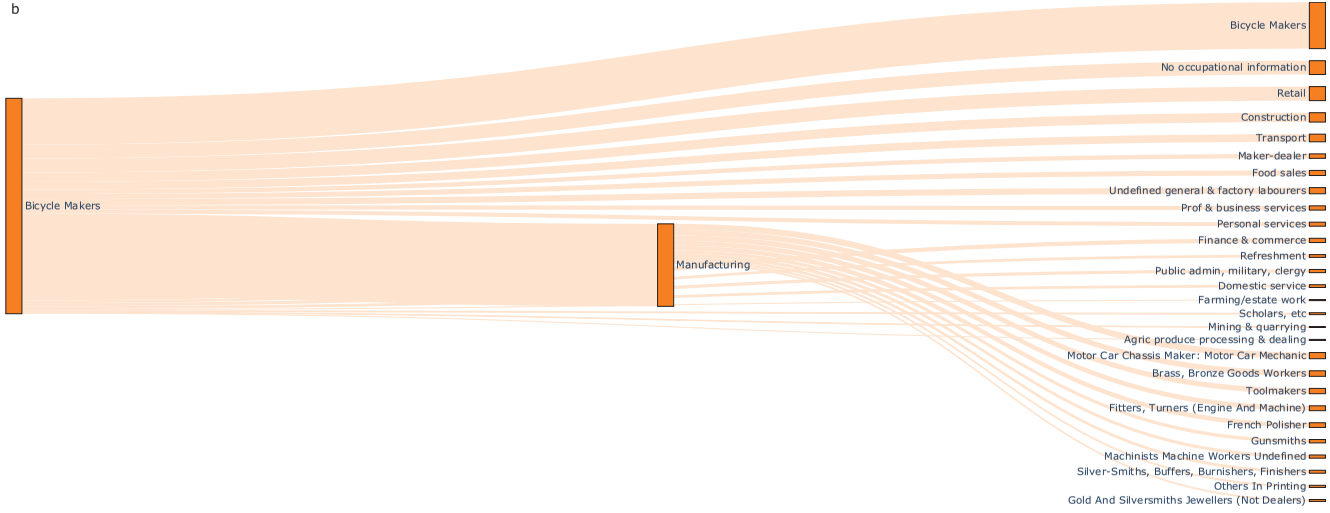


Figure 4.15b: (Cont.).

Ultimately, there are two main elements which drove this situation: geography and skills. The bicycle industry was largely geographically concentrated, despite its workshop structure. However, we find little evidence that it acted as a migratory pull by attracting workers from outside of its local area. One of the major explanations for the retention of workers is therefore geographic convenience: it was a sizable employer (particularly in Coventry) and was thus able to attract and retain labour. Importantly, this was in part dependent on the second element: skills. As most of the labour coming into the industry was new and without a developed pre-existing skillset, the bicycle manufacturing industry equipped those it employed with the skillset it required. That those who did leave the industry typically moved to other manufacturing jobs suggests that these skills were to some degree transferable, and that bicycle makers did not establish themselves as a distinct workforce lends weight to the idea that this skillset was not overly specialised. Reconstructing the large flows of labour into and out of the cycle trade suggests both that barriers to entry were low and that, at least for some workers, cycle making provided a bridge into more established, high-prestige metalworking occupations such as toolmaking and engine fitting customarily confined to formally apprenticed workers.

Conclusions

The bicycle manufacturing industry has often been regarded as one marked by transition. Born as a 'boom' industry, its rapid growth has been characterised as a success story of entrepreneurship and technological development. Its slower progress after 1900 was also tightly bound to technology: the rise of motor vehicle manufacturing – along with international competition and the saturation of the market – being key drivers, even if its continued international prominence mitigates against the notion of a true 'bust'. But technology still required humans, and what has previously been unclear is precisely how the industry cultivated and used the labour it required. We demonstrate that although bicycle making emerged within areas traditionally associated with manufacturing workers, it was not inherently these individuals taking advantage of these new bicycle making opportunities. Instead, the industry primarily recruited 'new' labour – which had not previously (or only recently) entered the labour force. Our *nominal linkage* technique therefore suggests it was the children of these workers who met the demand for labour that the bicycle boom created,

particularly those from declining trades such as watchmaking and ribbon making. The 'technological convergence' necessary to create the bicycle industry may well have been a product of the industrial composition of the Midlands, but it does not appear to have been particularly dependent on making use of the skillsets developed by its forerunner industries.

Our analysis also demonstrates the importance of *scale*. Whereas Coventry was dominated by bicycle manufacturing, Birmingham had a far more diverse manufacturing labour force, even though the two were comparable in terms of numbers employed in bicycle making for some years. The relative size and composition of different places therefore directly influenced the number of competing opportunities, and therefore inter-occupational flows of labour.

What is prominent in our analysis of both occupational origins and occupational destinations is the shifting and flowing nature of labour during this period. There were certainly core groups which remained employed as both bicycle makers (and metalworkers in general), but the ten-year snapshots provided by the census indicate that movement between related industries was consistently present in the labour force. Movement was generally confined to occupations within the sector (or at most, related sectors), which seemingly indicates some level of commonality between occupations, but the full implications of this for understanding skills acquisition and occupational life-course remain the subject of future research.

The power of our *nominal linkage* implementation lies in its ability to work at scale – both in terms of the size of the sample we are able to consider, and the information we are able to (reproducibly) utilise. However, the work presented here is only the tip of the proverbial iceberg. The scope and applications of this approach go far beyond the bicycle industry – or, indeed, occupational groupings in general. It is therefore important to separate the core functionality of our *nominal linkage* implementation from its intended use-case. The former relates to the task of matching individuals between two (or more) censuses. This is a technical exercise that requires careful consideration of the source materials, considerable testing and iterative refinement of various parameters. There are certainly further improvements which could be made in this regard – both for our specific use-case, and the application of this method elsewhere. However, while these are worthy of focus, debate and scrutiny, they are also distinct from the ultimate purpose of this task: the capacity to analyse groups of individuals over time.

It is this element of what we present here which offers the opportunity to fundamentally reshape research questions and engage with the census in new and insightful ways. The ability to match groups of individuals through multiple decennial censuses on the basis of any criteria (for example: geography, familial relationships, or immigration status) considerably expands the scope of what might reasonably be accomplished by researchers. Questions of community, social and economic mobility, family unit dynamics, and migration are particularly promising research areas – but ultimately, the flexibility of the method means that it could be applied to a broad array of topics which benefit from repeated observations of the same individuals.

It is also worth noting that although our approach is currently tailored to census records, the core principles are applicable to the linking of any structured population information. One area which might be particularly fruitful is the matching of other associated materials with census records (such as Civil Registrations of births, marriages and deaths), but utilising the census need not be a hard requirement. Potential applications will continue to grow as methods of data collection and extraction drawing on computational approaches develop – and with them, the opportunity to make use of data sources with looser structures and fuzzier information than the census. With this in mind, it is worth considering how the core elements of our approach might be generalised for other contexts.

Fundamentally, our approach is built upon the principles of interpretability and reproducibility. It is possible, if so desired, to take any match created by our implementation and understand exactly why the algorithm has linked those two records (and disregarded others). This is a conscious design decision, and one which is in direct contrast with the recent popularity of 'black box' techniques driven by the explosion of 'Artificial Intelligence'. We do not claim that this is the only way researchers should approach this problem – far from it, the incorporation of the latter provides promising avenues for future development. However, anyone working on these (or similar) problems must ensure that the nuances of the particular source materials are adequately incorporated and addressed. Without this, the richness of these materials risks being lost, and with them the subtleties necessary to accurately capture the complexities of historical societies. As we increasingly embrace working at scale, context and existing knowledge become even more vital.

There are undoubtedly improvements to be made, and the tuning of parameters to suit other research questions (in other words, the strictness of the criteria in Table 4.1) will certainly be necessary – although we deliberately adopt the position of favouring high-quality matches at the cost of a more limited sample size here, such an approach may not be appropriate for all research questions. An obvious area of improvement is the differentiation of potential ambiguous matches, which are excluded at present. Likewise, the development of more sophisticated and robust metrics for the measurement of errors will be necessary as work in this area develops. Nonetheless, the method we outline here provides a strong foundation for future work as well as new insights regarding the development of the bicycle industry in particular.

Appendix: occupational categories for the bicycle industry and metalworking trades

Although the general development of the bicycle industry is well known, there remains considerable uncertainty about the numbers of workers employed. Many manufacturers were small-scale workshops who contracted out specialised work such as enamelling or tin plating. As a consequence of these arrangements, some individuals explicitly mentioned 'bicycle' when reporting their occupations in the census, but many others returned general descriptions such as 'machine worker', 'metal turner', 'enameller' or other descriptions.

Census officials raised these concerns, noting that in 1881 and 1891 it was likely that 'the number was considerably understated at each enumeration, many men engaged [in the] industry being returned merely as metal workers of various kinds'.[44] Similarly, in 1901 census aggregations failed to differentiate between cycle and motorcar makers, whereas these were enumerated separately in 1911, meaning that a direct comparison of the two years would underestimate growth in cycle employment across the decade.[45] Table 4.2 shows a comparison between the number of bicycle makers and dealers as reported in the contemporary printed census reports, and also identifiable in I-CeM through its occupational coding scheme. The totals broadly correspond for 1881, 1891 and 1901, but not in 1911, when around 10,000 fewer workers are classified as bicycle makers in I-CeM than that year's census report. Most likely, this is because individuals working in the bicycle industry have not been coded to the associated

Table 4.2: Number of bicycle workers, 1881–1911[46]

Year	Census Reports			I-CeM					
	England and Wales						Scotland		
	M	F	All	M	F	All[b]	M	F	All[b]
1881	1,059	13	1,072	891	58	949	19	0	19
1891	10,965	559	11,524	8,777	457	9,635	101	0	101
1901[a]	28,537	2,929	31,466	23,965	2,456	27,393	950	55	1,005
1911	29,063	6,328	35,391	21,983	3,153	25,138	-	-	-

Notes:

I-CeM figures are the number of individuals coded to Occode '362' (Bicycle Makers, Bicycle Repairers) and '374' (Bicycle dealers). The '374' I-CeM coding erroneously included many other occupations, so only individuals with the word 'cycle' in their occupation title have been included from the '374' Occode subset. Census Report figures refer to 'bicycle and tricycle makers and dealers'.

[a] Bicycle and motor vehicle manufacture were listed combined under 'Cycle and Motor Manufacture' in 1901, but were separated out again in 1911. The 1911 Census Report noted that 'doubtless at that date [1901], the great majority of the workers [. . .] were engaged in cycle manufacture'.

[b] All totals for I-CeM include people recorded as male, female and unknown. Any discrepancy between the total of the Male and Female columns and the All column is the number of people whose sex is 'unknown' in I-CeM.

Occode, '362 – Bicycle Makers, Bicycle Repairers' due to variations in the way in which occupations relating to the bicycle industry were recorded by the census. As we consider all possible options for a match, this does not prevent us from tracking individuals between censuses – although it may have a limited influence on the size of our sample.

There is a further Occode relating to the bicycle industry present in I-CeM, '374 – Bicycle Dealers' – although this is comparatively much smaller, and more focused on the sale of bicycles than their manufacture. In addition, our evaluation of the actual occupations returned by individuals coded to either 362 or 374 shows that 374 includes many individuals who quite obviously did not work as bicycle dealers. For example, among those coded to 374 in 1891, we find four 'Manager of Living Stables' and four 'Manager of Dairy Milk'. Overall, only 344 of the 2,701 coded to 'Bicycle Dealers' in England and Wales in 1891 listed 'bicycle' or 'cycle' in their occupation; by contrast, 97 per cent of the 9,291 persons coded to 362 in the same year had occupations referring to cycling. We therefore adopt those assigned to Occode 362 as our sample of bicycle makers.

Our 'composite metalworker' classification also draws on the Occode coding scheme, but, as noted, it is a composite of fifty-nine categories excluding bicycle making – as given in Table 4.3. Although we examine trends for those employed in this grouping, it is primarily used as a control group for the bicycle making sample to enable us to assess whether the new industry was more or less open to recruitment from outside the sector than metalworking as a whole.

There are several reasons for creating a composite occupational grouping for this purpose, rather than using either a single occupation or broader sector as a comparator. First, we require our control group to demonstrate the general trend in this area of the economy over time, but also to be resistant to large swings as a consequence of booms or busts in particular industries. An analysis comparing bicycle makers to 'motor car body makers', for example, would be informative as a direct comparison of the two related occupations, but both were subject to rapid changes in the size and nature of their workforces. Adopting a panel of occupations provides a longer-term trajectory of individuals working in this general area. Conversely, a sectoral comparator (such as 'Manufacturing') would give a good impression of longer-term economic trends but also incorporate a variety of occupations not directly comparable. Our composite grouping

Table 4.3: Occupations in our 'composite metalworker' classification[47]

Occode	Description (standardised)
253	Patternmakers (Engine and Machine Making, Undefined – Not Textile)
254	Patternmakers (Spinning Weaving Machinery Making)
255	Patternmakers (Agricultural Machine and Implement Making)
256	Patternmakers (Domestic Machinery Making)
257	Patternmakers (Undefined)
258	Millwright
259	Ironfounder – Moulders, Core Makers, Fetter, Cupola, Overn-Man, Foundry Labourer, Other and Undefined
260	Brass Founders
261	Farriers
262	Blacksmiths
263	Fitters, Turners (Engine and Machine)
264	Colliery Fitters
265	Railway – Signal Switch Turntable Fitters
266	Fitters, Turners (Engine and Machine) Labourers
267	Brass Finishers
268	Coppersmiths (Various)
273	Undefined Engine and Machine Making (Not Textile)
275	Boiler Makers
276	Other Workers (Spinning Weaving Machinery Making)
277	Other Workers in Textile Machinery Fittings
278	Engine And Machine Makers (Undefined – Not Textile)
279	Domestic Machinery Makers
280	Agricultural Machine and Implement Makers
281	Railway – Signal Switch Turntable Makers
282	Riveters Riggers Metal Plate Workers (Not Ships, Textiles, Boilers)
283	Technicians (Unspecified)
284	Electrical Cable Makers

285	Electrical Lamp Makers
286	Other Electrical Apparatus Makers
289	Toolmakers (Agricultural Machine and Implement Making)
290	Toolmakers
291	Domestic Implement Makers – Toolmakers
301	Gunsmiths
315	Lock, Key Makers
316	Gas Fittings Makers
317	Lamp, Lantern, Candlestick – Maker
318	Silversmiths, Buffers, Burnishers, Finishers
319	Pewter White Metal Plated Ware Manufacture
320	Metal Refiners Workers
321	Tinplate Goods Manufacture
322	Copper Workers
323	Leaden Goods Manufacture (Various)
324	Zinc Goods Workers
325	Brass, Bronze Implement Makers
326	Brass, Bronze Goods Workers
343	Whitesmiths
345	Ship Boat Platers Riveters
346	Metal Ships – Workers in Iron
347	Fitters (Ships)
349	Shipwrights – Metal Ships
360	Railway – Coach, Wagon Maker
361	Tram Car Maker
363	Motor Car Chassis Maker: Motor Car Mechanic
364	Motor Car Body Maker
365	Coach, Carriage – Maker
382	Gold and Silversmiths Jewellers (Not Dealers)
384	Gold and Silver Beaters
385	Other Workers Gold Silver Jewellery
768	Artisans Mechanics

Note: Original spellings of overn-man (oven-man) and fetter (fettler) are retained.

therefore allows us to compare trends in the bicycle manufacturing industry to a panel of similar workers over the period of study.

The composite metalworker group includes all metalworking occupations except the highly specialist and localised cutlery-making industry while excluding workers explicitly identified as labourers.[48] In addition, metal manufacturing jobs such as iron and steel-making are excluded because craft control was much weaker in these trades and the workshop system that dominated light metalworking was unknown.[49]

Notes

1 Authorship statement: This chapter presents work undertaken as part of the 'Occupations' work package on the Living with Machines project. The general scope of this work was conceived by Griffin and Lawrence. The original methodology was devised by Rhodes and Griffin, with later input from Lawrence, Hobson, Solomon, Lo and Wilson. The initial software implementation was created by Rhodes, and further developed in collaboration with Hobson, Lo and Solomon. The secure data storage environment, which enabled access to these data, was maintained by Hobson. Research questions were developed collaboratively, with input from all authors. The original text of this chapter was written by Solomon, Lawrence, Rhodes and Wilson, with input from Griffin and Hobson. Data analysis was designed and implemented by Solomon and Rhodes, with input from all authors (but particularly Lawrence). Historical analysis was led by Lawrence, with input from Solomon, Rhodes and Wilson. Visualisations were created by Solomon. The final version was revised by Solomon, Lawrence, Rhodes and Wilson.

2 For example: Emma Griffin, *Bread Winner: An Intimate History of the Victorian Economy* (Yale University Press, 2020).

3 Jerome De Groot, 'The Genealogy Boom: Inheritance, Family History, and the Popular Historical Imagination', in *The Impact of History? Histories at the Beginning of the Twenty-First Century*, ed. Pedro Ramos Pinto and Bertrand Taithe (Routledge, 2015), 21–33.

4 Our choice of topic was influenced by the absence of data for 1871 from the public census dataset which currently hampers longitudinal analysis of demographic change in nineteenth-century Britain.

5 David V. Herlihy, *Bicycle: The History* (Yale University Press, 2004), 235–50.

6 A.E. Harrison, 'The Origins and Growth of the UK Cycle Industry to 1900', *The Journal of Transport History*, 6, no. 1 (1985): 43–5; Anthony Edward Harrison, 'Growth, Entrepreneurship and Capital Formation in the United Kingdom's Cycle and Related Industries, 1870–1914', (Unpublished PhD thesis, University of York, 1977), 5.

7 Damien Kimberley, *Coventry's Bicycle Heritage* (The History Press, 2015), 86, 142, 153.

8 Wiener's declinist thesis was later challenged by Chaloner. Martin J. Wiener, *English Culture and the Decline of the Industrial Spirit, 1850–1980* (Cambridge University Press, 1981); William Henry Chaloner, 'Was There a Decline of the Industrial Spirit in Britain 1850–1939?', *Transactions of the Newcomen Society*, 55 (1985 for 1983–4): 214; Harrison, 'Origins and Growth'.

9 D.W. Thoms and T. Donnelly, 'Coventry's Industrial Economy, 1880–1980', in *Life and Labour in a 20th Century City: The Experience of Coventry*, ed. Bill Lancaster and Tony Mason (Cryfield Press [c.1980]), 12–21.

10 Bill Lancaster and Tony Mason, eds, *Life and Labour in a 20th Century City: The Experience of Coventry* (Cryfield Press, n.d. [c.1980]); David McGrory, *A Century of Coventry: Events, People and Places Over the 20th Century* (Sutton Publishing, 2007).

11 Pat Hudson, ed., *Regions and Industries: A Perspective on the Industrial Revolution in Britain* (Cambridge University Press, 1989).

12 S. Kenny, 'Sub-regional Specialization in the Lancashire Cotton Industry, 1884–1914: A Study in Organizational and Locational Change', *Journal of Historical Geography*, 8, no. 1 (1982): 41–63; Frederick J. Glover, 'The Rise of the Heavy Woollen Trade of the West Riding of Yorkshire in the Nineteenth Century', *Business History*, 4, no. 1 (1961): 1–21.

13 'Willenhall: The Town of Locks and Keys', Willenhall History Society, accessed 24 February 2026, www.willenhallhistory.co.uk/society/locksandkeys.htm, and 'A History of Wednesfield, 4: the Nineteenth Century', accessed 24 February 2026, www.historywebsite.co.uk/articles/Wednesfield/nineteenth.htm.

14 West Midland Group, *Conurbation: A Planning Survey of Birmingham and the Black Country* (Architectural Press, 1948), 36–7, 111–21; Thoms and Donnelly, 'Coventry's Industrial Economy'.

15 Alan Fox, 'Industrial Relations in Nineteenth-Century Birmingham', *Oxford Economic Papers*, 7, no. 1 (1955): 57–70; Martin J. Daunton, *Wealth and Welfare: An Economic and Social History of Britain, 1851–1951* (Oxford University Press, 2007), 82, 84–7.

16 Nathan Rosenberg, 'Technological Change in the Machine Tool Industry, 1840–1910', *The Journal of Economic History*, 23, no. 4 (1963): 423.

17 Harrison, 'Origins and Growth', 50; Glen Norcliffe, 'G-COT: The Geographical Construction of Technology', *Science, Technology, & Human Values* 34, no. 4 (2009): 461–5.

18 Harrison, 'Origins and Growth', 50–1.

19 Herlihy, *Bicycle*, 163, 184–5.

20 Norcliffe, 'G-COT', 470–1.

21 Kimberley, *Coventry's Bicycle Heritage*, 90–91, 124, 138–9, 160.

22 Paul Rosen, *Framing Production: Technology, Culture and Change in the British Bicycle Industry* (MIT Press, 2002), 1–40; Harrison, 'Origins and Growth', 50.

23 Harrison, 'Origins and Growth', 68n11.

24 Herlihy, *Bicycle*, 283–5.

25 Herlihy, *Bicycle*, 310–15.

26 Thoms and Donnelly, 'Coventry's Industrial Economy', 13, 20; Kenneth Richardson, *Twentieth-Century Coventry* (Macmillan, 1972), 33–7.

27 Herlihy, *Bicycle*, 315–16. On the continued growth of cycle employment in Britain after 1901 see, *Census of England and Wales, 1911*, vol. 10, pt 1, *Occupations and Industries* [Cd. 7018] (HMSO, London, 1914), lxxvi–lxxvii.

28 Bill Lancaster, 'Who's a Real Coventry Kid? Migration into Twentieth Century Coventry', in *Life and Labour*, ed. Lancaster and Mason, 59–60.

29 I-CeM has been available since 2014 with data for 1851–1911, it was updated in 2022 to correct errors, and a new version including 1921 census data was released in 2024. The 1921 data were added after the Living with Machines project ended and so this chapter only uses data up to 1911: K. Schurer and E. Higgs (2020), *Integrated Census Microdata (I-CeM), 1851–1911* [data collection]. UK Data Service. SN: 7481, http://doi.org/10.5255/UKDA-SN-7481-2; K. Schurer and E. Higgs (2020), *Integrated Census Microdata (I-CeM) Names and Addresses, 1851–1911: Special Licence Access* [data collection], 2nd ed., UK Data Service. SN: 7856, http://doi.org/10.5255/UKDA-SN-7856-2.

30 Occupational coding ('Occode') is part of I-CeM, but a look-up version is available in Kevin Schurer, 'I-CeM Lookup Table – Occupation Code'. Apollo – University of Cambridge Repository, 2024, http://doi.org/10.17863/CAM.106514. We identify the individuals living (and, most likely, working) in particular locations using the parish recorded by the census, and the urban classification coding scheme created by the British Business Census of Entrepreneurs project: Harry Smith and Robert J. Bennett, 'Research Data Supporting WP 6: Urban-Rural Classification Using Census Data, 1851–1911', Apollo – University of Cambridge Repository, 2019, http://doi.org/10.17863/CAM.44400. For further details on our use of I-CeM, see Appendix.

31 While this is unfortunate, the Scottish bicycle industry was relatively small (see Table 4.2) and thus is unlikely to alter the overall results.

32 By comparison, we identify 538,456 composite metalworkers in 1881; 679,375 in 1891; 846,842 in 1901; and 991,702 in 1911.

33 The county boundaries in Figure 4.1 and Figure 4.2 are derived from RSD boundary files and Lookup Dictionaries supplied directly by Joe Day at the University of Bristol and Alice Reid at the University of Cambridge. J.D. Day, 'Registration Sub-district Boundaries for England and Wales 1851–1911' (2016). This dataset was created by the 'Atlas of Victorian Fertility Decline' project (PI: A.M. Reid) with funding from the ESRC (ES/L015463/1). Interactive versions of all figures in this chapter are included in the open access edition of the book on the University of London Press instance of the Manifold platform: https://read.uolpress.co.uk/projects/living-with-machines.

34 For work on the United States, see M. Bailey, P.Z. Lin, A.R.S. Mohammed, P. Mohnen, J. Murray, M. Zhang and A. Prettyman, 'The Creation of LIFE-M: The

Longitudinal, Intergenerational Family Electronic Micro-Database Project', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 56, no. 3 (2023): 138–59. For other areas, see Ö. Akgün, A. Dearle, G. Kirby, E. Garrett, T. Dalton, P. Christen, C. Dibben and L. Williamson, 'Linking Scottish Vital Event Records Using Family Groups', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53, no. 2 (2019): 130–46; S. Galli, D. Theodoridis and K. Rönnbäck, 'Reconstructing a Slave Society: Building the DWI Panel, 1760–1914', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57, no. 3 (2024): 163–84.

35 Kasey Buckles, Adrian Haws, Joseph Price and Haley E.B. Wilbert, 'Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project', National Bureau of Economic Research Working Paper Series, 2023, Working Paper 31671. For pioneering work on historic record linkage, see, for example, Jeremy Atack, Fred Bateman and Mary Eschelbach Gregson, 'Matchmaker, Matchmaker, Make Me a Match', *Historical Methods* 25, no. 2 (1992): 53–65; J.P. Ferrie, 'A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules', *Historical Methods* 29, no. 4 (1996): 141–56; and the work arising from the Minnesota Population Center, for example: Steven Ruggles, 'Linking Historical Censuses: a New Approach', *History and Computing* 14, nos. 1–2 (2002): 213–24. Also see, R. Abramitzky, R. Mill and S. Pérez, 'Linking Individuals Across Historical Sources: A Fully Automated Approach', *Historical Methods* 53, no. 2 (2020): 94–111; Catherine G. Massey, Katie R. Genadek, J. Trent Alexander, Todd K. Gardner and Amy O'Hara, 'Linking the 1940 U.S. Census with Modern Data', *Historical Methods* 51, no. 4 (2018): 246–57; Jonas Helgertz, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles and Catherine A. Fitch, 'A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel', *Historical Methods* 55, no. 1 (2022): 12–29; Ron Goeken, Lap Huynh, T. A. Lynch and Rebecca Vick, 'New Methods of Census Record Linking', *Historical Methods*, 44, no. 1 (2011): 7–14.

36 For notable exceptions, see Gill Newton and Robert Bennett, *Record-Linkage of Entrepreneurs in the England and Wales Censuses 1851–91 Using BBCE and I-CeM* (2020), http://doi.org/10.17863/CAM.50178; Z. Zhu, 'Like Father Like Son? Intergenerational Immobility in England, 1851–1911', *The Journal of Economic History* 84, no. 4 (2024): 1143–73; R. Thomas, 'The Victorian Intergenerational Migrant Cohort: Socioeconomic Outcomes, 1851–1901' (Unpublished MPhil thesis, University of Oxford, 2018).

37 We use fuzzy string matching to determine similarity of names and places. Two entries must score ≥ 0.85 (1 being an exact match) on Jaro-Winkler distance to be considered a match. For information on the implementation and use of Jaro-Winkler, see Matthew A. Jaro, 'Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida', *Journal of the American Statistical Association* 84, no. 406 (1989): 414–20, http://doi.org/10.1080/01621459.1989.10478785; William E. Winkler, 'String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage', in American Statistical Association, *Proceedings of the Section on Survey Research Methods* (1990): 354–359. We employ the implementation found in: J. De Bruin,

'Python Record Linkage Toolkit: A Toolkit for Record Linkage and Duplicate Detection in Python. v0.14', Zenodo, 1 December 2019, http://doi.org/10.5281/zenodo.3559043.

38 Although this could still overstate the number of 'true' matches, due to the potential for false positives that cannot be identified without extensive reconstruction of life histories for each matched individual that cross reference sources beyond the census.

39 For clarity, we adopt a defined structure and colour scheme in each figure: the central (blue) bar indicates the number of people employed in that occupational grouping in each year, the right-most (orange) bar shows the number of those individuals which we link forwards to the next census, and the left-most (green) bar shows the number of individuals we link to the previous census. In following figures, the green/orange colour scheme is preserved.

40 Rosenberg, 'Technological Change in the Machine Tool Industry'.

41 For sectors, we use a slightly modified version of Bennett et al.'s 'Economic Active: 17 Sector framework', which map onto the I-CeM Occodes. Our modifications are: (1) to split out bicycle makers or composite metalworkers (whichever is the focus) from their sector, and (2) we rename the 'Persons of property' sector to 'Scholars, etc', to make it clearer that people we identify in this sector are almost universally classified as 'Scholars' in the previous census: Robert Bennett, Harry Smith, C. Van Lieshout and G.H. Newton, 'Sector and Occupation Classification Data Download for Business Sectors, Occupations and Aggregations of Census Data 1851–1911'. Apollo – University of Cambridge Repository (2018), http://doi.org/10.17863/CAM.26433.

42 Please note that the occupation-level data appears only in the online version of Figures 4.9–12.

43 'The City of Coventry: Crafts and Industries, Modern Industry and Trade', in Victoria County History, *A History of the County of Warwick*, vol. 8, *The City of Coventry and Borough of Warwick*, ed. W.B. Stephens (Oxford University Press for the Institute of Historical Research, 1969), *British History Online*, accessed 22 July 2024, https://www.british-history.ac.uk/vch/warks/vol8/pp162-189. See also Alun C. Davies, *The Rise and Decline of England's Watchmaking Industry, 1550–1930* (Routledge, 2022); Richardson, *Twentieth-Century Coventry*, 14–15, 33.

44 *Census of England and Wales, 1891,* General Report, vol. 4, 48, https://www.visionofbritain.org.uk/census/EW1891GEN/7 (accessed 29 October 2022).

45 *Census of England and Wales, 1911*, vol. 10, pt 1, *Occupations and Industries*, [Cd. 7018] (HMSO, London, 1914), lxxvi–lxxvii.

46 Sources: Census of England and Wales (43 & 44 Vict. c. 37). 1881. Volume III, Ages, Condition as to Marriage, Occupations, and Birthplaces of the People (Eyre and Spottiswoode, 1883), xiii, https://archive.org/details/cu31924063571123/page/n25/mode/2up (accessed 29 October 2022); Population. Age, marriage, occupation, birth-place, England and Wales, vol. 3, 1891, xxvii. Census of England and Wales (63 Vict. c. 4), 1901, General Report with Appendices (Darling & Son, 1904), 111. https://archive.org/details/censusenglandan01offigoog

/page/n126/mode/2up (accessed 29 October 2022); Census of England and Wales, 1911 (10 Edward 7 and 1 George 5, Ch. 27). Volume X, Occupations and Industries, Part I (Athlone Printing Works Company, 1914), lxxvi. https://archive.org/details/b32183185_0005/page/n79/mode/2up (accessed 29 October 2022); 1891 Census Report data as cited in A.E. Harrison, Growth, Entrepreneurship and Capital Formation in the United Kingdom's Cycle and Related Industries, 1870–1914, PhD thesis, University of York, 1977, 4.

47 Source data now under Kevin Schurer, 'I-CeM Lookup Table – Occupation Code', Apollo – University of Cambridge Repository (2024). http://doi.org/10.17863/CAM.106514.

48 Sally-Ann Taylor, 'Tradition and Change: The Sheffield Cutlery Trades, 1870–1914' (Unpublished PhD Thesis, University of Sheffield, 1988).

49 Michael Savage, *The Dynamics of Working-Class Politics: The Labour Movement in Preston, 1880–1940* (Cambridge University Press, 1987), 43, 48–9.

Chapter 5

Analysing the language of mechanisation in nineteenth-century British newspapers

Barbara McGillivray, Nilo Pedrazzini, Arianna Ciula, Jon Lawrence, Tiffany Ong, Mia Ridge and Miguel Vieira[1]

Introduction: mechanisation in nineteenth-century Britain and the English lexicon

Mechanisation in nineteenth-century Britain was far more than a series of technological advancements, and it deeply influenced and shaped the nation's socio-cultural fabric. The effects of these shifts had strong repercussions on various aspects of Late Modern English.[2] The English lexicon evolved to accommodate a myriad of new terms, expressions and nuances associated with the growing industrial landscape and cityscape, reflecting the profound societal shifts, and the perceptions, cultural ethos and anxieties surrounding technology at the time.[3] We present the results of our linguistic analysis of nineteenth-century newspaper sources to trace the semantic evolution of some English words related to the domain of mechanisation. The main argument of this chapter is that the language of mechanisation, as reflected in the multifaceted discourse of nineteenth-century newspapers, offers a unique perspective on the societal perceptions, aspirations and anxieties that accompanied the sweeping technological and cultural changes of the period. Employing computational methods complementing the historical and linguistic analysis, we show how these tools, enhanced by data visualisations, reveal patterns of lexical

change, shedding light on the profound interplay between language and society during a time of transformative change.

The nineteenth century marked a significant period of transformation in education and literacy, characterised by gradual but consistent improvements in educational opportunities and by the proliferation of printed media. Prior to the introduction of compulsory elementary education in England in 1870, education was fragmented across various institutions, and adult education played a crucial role in expanding literacy rates.[4] The rise of popular novels and newspapers, which saw a rapid increase in number and distribution during this period, exerted a profound influence on readers and contributed to the spread of literacy. The mid-nineteenth century witnessed a remarkable surge in the production of books focused on literary instruction and grammar, reflecting societal developments and the growing public interest in education. A concerted effort was made for the first time to educate the labouring poor in rural areas and industrial towns, facilitated by various economic developments such as the introduction of fast and affordable printing techniques, the repeal of taxes on newspapers and paper, and the establishment of circulating libraries and working-men's reading clubs.[5] The expansion of literacy had significant implications for the dissemination of Standard English in written form and contributed to the democratisation of styles in literature and journalism. Moreover, the standardisation and codification of English accelerated in the nineteenth century, with dictionaries and linguistic scholars endeavouring to capture the breadth and depth of the language. Technological advances, social changes, geographical expansion, and contact with other languages led to a particularly dynamic phase in the history of the English language, during which the vocabulary underwent rapid changes.[6]

One of the salient features of the nineteenth-century English lexicon is the influx of new words stemming from the rapid advancements in science, technology and industry, such as 'telegraph' or 'telephone'. Another significant aspect of the nineteenth-century English lexicon is the semantic evolution and expansion of existing words via the acquisition of new meanings in relation to their application to emerging technical advancements and shifts in meaning as new connotations or evaluations became associated with them.[7]

Several studies have analysed the portrayal of machines and technology in Victorian fiction and culture, focusing on literary landscapes.[8] While

some attention has been devoted to the role of periodicals, the historical linguistic analysis of newspaper sources is still underexplored.[9] As non-canonical sources, newspapers offer a kaleidoscopic view of the era, capturing the many voices that shaped the discourse of the time and offering an opportunity to uncover perspectives that often remain marginalised in traditional historical narratives. This chapter explores the semantic changes affecting the lexical field of mechanisation in the English lexicon during the nineteenth century and relates them to the profound technological, cultural and social transformations of the time. We base our analysis on a portion of the British Newspaper Archive (BNA), which (at the time of writing) contains over 80 million pages of newspapers dating from the 1700s. Large digital corpora provide researchers with rich and extensive datasets for analysing lexical and semantic changes over time, identifying patterns and uncovering previously unnoticed trends or phenomena.[10] Additionally, advanced computational methods from natural language processing (NLP) allow us to perform quantitative assessments of linguistic data. Last but not least, visualisation techniques with inspectable code can offer opportunities to explore these data from a variety of perspectives, aiding interpretability, verifiability and inspiring further analysis.

This research took place as part of the 'Language of Mechanisation' strand within the Living with Machines project, and involved an interdisciplinary team of linguists, historians, research software engineers, curators and other library professionals, and research software engineers. In the spirit of 'radical collaboration', the research also involved hundreds of members of the public, whose ongoing feedback made possible the groundbreaking design of a crowdsourcing task for the semantic annotation of historical texts, described below. Our interdisciplinary analysis sheds light on how words related to mechanisation evolved, often in unexpected and nuanced ways. Compared to traditional sources such as the *Oxford English Dictionary* (*OED*), which provides a selective and sometimes biased record of language change, our analysis of digitised newspaper archives offers a complementary, more granular view, particularly within the specific communicative contexts of historical newspapers.[11] The chapter also contributes to corpus linguistics methodology thanks to its methods for tracking semantic change in large, messy historical datasets. By combining computational methods with historical and linguistic

analysis, we offer a robust framework for studying lexical evolution in historical corpora, which can be generalised to other periods and datasets. Moreover, our approach to mapping lexical trends and quantifying shifts in word meaning opens up the possibility of identifying significant semantic changes without preselecting terms, enabling researchers to detect which words or clusters undergo the most notable transformations.

The next section describes the work we have done in developing methods for the automatic detection of lexical semantic change, that is, the phenomenon by which the meaning of words changed over time. We were able to identify when each word started being used in different contexts and with different meanings by training diachronic word embedding models on digitised nineteenth-century British newspaper articles. Word embeddings are numerical representations of words that reflect their evolving meanings and usage patterns across different time periods and are widely used in computational linguistics and NLP to analyse changes in words' meaning over time.[12] We also report on our work looking at diatopic (geographic) variation in the lexicon of mechanisation – that is, how the words were used differently in different regions covered by our newspaper corpus – by leveraging work outlined in Chapter 1, on the Environmental Scan. This describes at more length the creation of enhanced metadata for the newspaper titles in the project's datasets from nineteenth-century press directories digitised as part of the project. Our results from experiments using this enhanced metadata highlighted certain geographical areas where words changed earlier compared to other regions.

Moving to a smaller scale and human annotation methods, the 'Designing sense disambiguation tasks for crowdsourcing volunteers' section describes our work designing, implementing and optimising the semantic annotation of a portion of our newspaper corpus via voluntary crowdsourcing. This section introduces novel 'close reading' citizen science tasks we designed, inviting volunteers to read texts and select the most appropriate definition for words appearing in specific contexts. Datasets containing high-quality human semantic annotations of historical texts, where specific words are annotated with their meaning in context, are very rare. This task not only engaged members of the public with research on nineteenth-century newspapers but also allowed us to publish the first datasets annotated at the level of word meaning for historical

English.[13] These datasets are relevant to historical linguists working on the evolution of the English lexicon in the nineteenth century. They also provided insights into the evolving language associated with specific types of machines during this transformative period.

The final section describes our work in visualising and analysing the crowdsourced annotation data. We developed an interactive Notebook which allows the user to explore the dataset of source texts and annotations. The Notebook also makes available for examination the code used to process and visualise the data. The Notebook allowed us to examine patterns, trends and variations in the lexical semantics related to mechanisation in the newspaper corpus. The Notebook interface is designed to enable users to filter and sort the data based on various criteria, such as geographical regions, specific lexical items or political leaning of the newspaper titles. These functionalities allow for targeted investigations and comparisons, offering opportunities to develop a nuanced understanding of how language evolved in response to mechanisation across different contexts. The development of this interactive tool enhances the accessibility and usability of the crowdsourced annotation data and deepens our analysis of semantic change associated with mechanisation.

Automatic detection of meaning change in the lexicon of mechanisation

Historians have long recognised that handcraft skills and mechanisation continued to exist side-by-side throughout Britain's gradual rise to industrial pre-eminence.[14] The study of linguistic variation across registers, social class and geography, in addition to change over time, requires a deep engagement with historical scholarship. In our investigation, we sought to explore the evolution of word meanings over the nineteenth century by focusing on words which previous historical and lexicological scholarship had identified as key vocabulary related to mechanisation and as having registered significant meaning change across the period.

Our main point of departure was Manfred Görlach's discussion of language change in nineteenth-century English.[15] We selected words mentioned as examples of meaning change that occurred as a result of mechanisation, specifically 'bulb', 'coach', 'gear', 'matches', 'railway', 'stamp', 'stock', 'traffic', 'train', 'trade' and 'wheel'.

The corpus of historical British newspapers used to carry out our computational analysis comprises around 4.6 billion words and spans the period between 1801 and 1920. It includes titles digitised by the British Library through an internal project that focused on full runs of metropolitan newspapers with a wider circulation (i.e. newspapers published in London and distributed beyond London, noted as a neglected area for newspaper digitisation), where some of the volumes for that title were in a preservation condition internally called 'unfit' (e.g. with a broken spine, crumbling pages, etc.) (around 2.3 billion words) and the titles specifically selected and digitised from the British Library newspaper collection for the Living with Machines project (a further 2.3 billion words).[16]

As the flowchart in Figure 5.1 shows, we split the newspaper corpus into decade-long 'time slices', namely separate subcorpora containing only articles published within a certain year range, which could be achieved thanks to the associated newspaper metadata, including the year of publication. We decided to work with ten-year time slices to be able to track potential changes in the meaning of words from one decade (e.g. the 1840s, i.e. 1840–1849) to the next (e.g. the 1850s). Because of the large size of the corpus, we only applied minimal pre-processing, that is, lowercasing and punctuation removal, but no lemmatisation (i.e. standardising of words to their base form, e.g. 'making' and 'made' to 'make'). We then trained word vector representations (or 'word embeddings') with Word2Vec for each decade.[17] Word embeddings represent words in a numerical form as points in a multidimensional space. The location of these points is determined by the linguistic contexts in which the words appear. Words that frequently appear in similar contexts (e.g. 'coffee' and 'tea') are positioned in close proximity because they share usage patterns (i.e. drinking, breakfast, hot beverages, etc.). This spatial arrangement of words allows for the modelling of the vocabulary as a landscape where semantic relationships can be visually and mathematically discerned. For instance, analogies between words can be captured by adding or subtracting the vectors of the words in question, as in the now classic example 'king' - 'man' + 'woman' = 'queen'. By visualising the respective points, the semantic relations between these words will also be reflected in their position relative to each other, as in Figures 5.2a–b.

Following an established practice in semantic change detection research,[18] we ensured that the vectors for the same words across decades

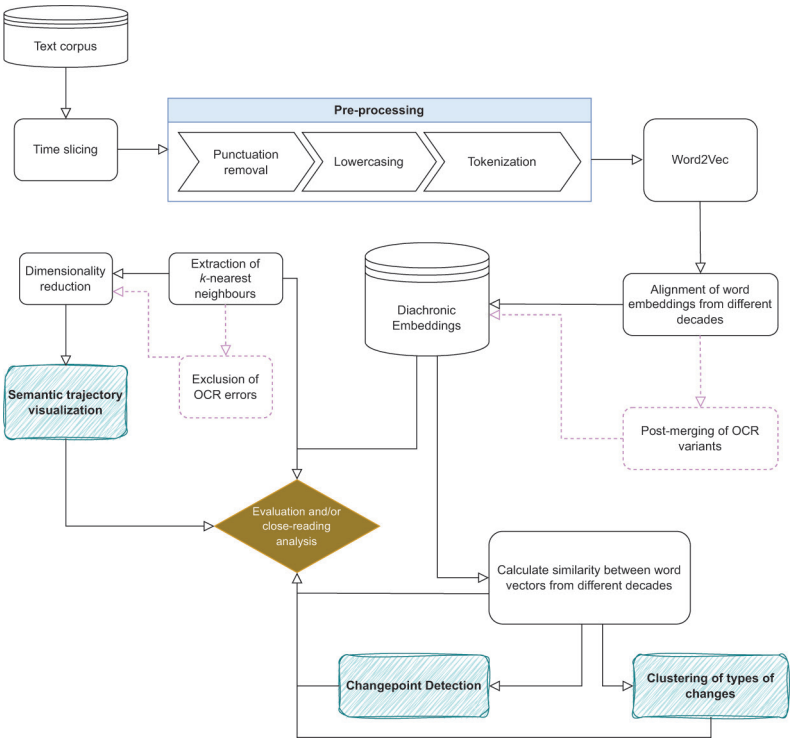


Figure 5.1: Flowchart of the methods for the automatic detection of meaning change with diachronic word embeddings, from pre-processing to time-series analysis. Steps involving the post-correction of Optical Character Recognition (OCR) errors are indicated by dashed boxes, since they are optional as far as the training and visualisation of diachronic word embeddings are concerned. Sketched edges indicate automatic methods for the analysis of diachronic embeddings. Credit: Nilo Pedrazzini.

were comparable by 'aligning' all the embedding models to the model for the most recent decade available (1910s).[19] The result was one embedding model for each decade in the nineteenth century, where every word occurring in the subcorpus for the respective decade is represented by a vector (i.e. a sequence of numbers which can be visualised as a point in a geometrical space), as shown in the schematised entries taken from the 1860s and 1870s models for 'England', 'house', 'machine' and 'train' in Figure 5.3.[20]

At this point, we have aligned, hence comparable, representation of the same words across decades, namely, *diachronic* word embeddings.

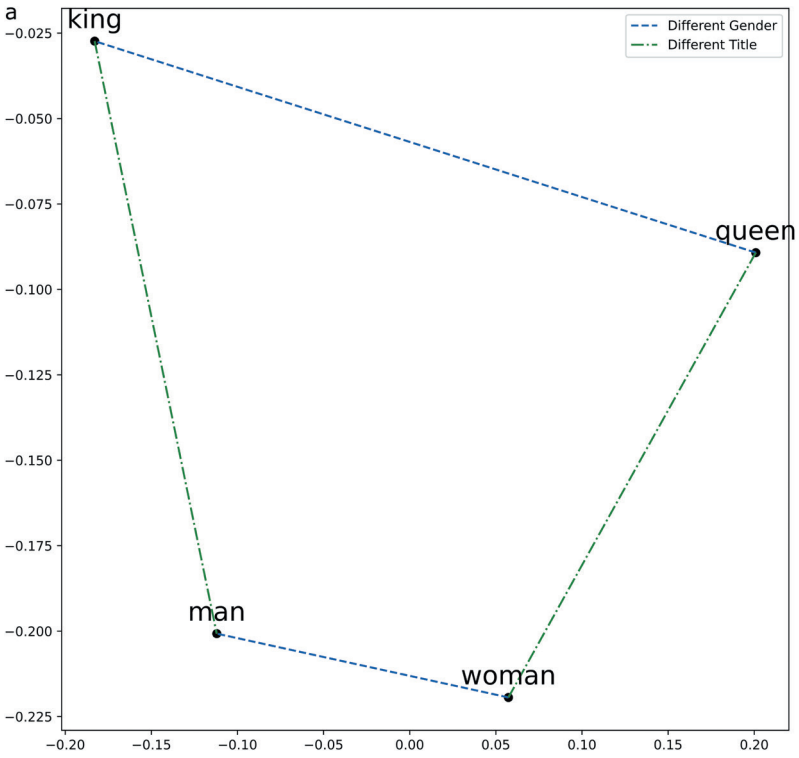


Figure 5.2a: Example of word embeddings as points in space, produced from the word embedding model trained on the 1880s newspaper subcorpus. The coloured lines represent a possible semantic interpretation of the position and relative distance between each point. Credit: Nilo Pedrazzini.

Working with digitised historical texts, however, often means dealing with OCR errors, inevitably created during the newspaper digitisation process. A possible, though computationally expensive, way of partially counteracting the noise created in the embeddings trained on OCR'd texts is to attempt error post-correction. Doing so directly on the text sources would be computationally infeasible for very large data such as our newspaper corpus, so we opted to update the trained embeddings of spelling variants likely due to OCR errors, by averaging the respective vectors and keeping only the likely correct entry in our corpus.[21] For example, the method found the entries 'eugland', 'ngland' and 'rngland' to be orthographic variants of the entry 'england' due to OCR errors. Therefore, the respective vector will be merged with the vector for 'england' by averaging them:

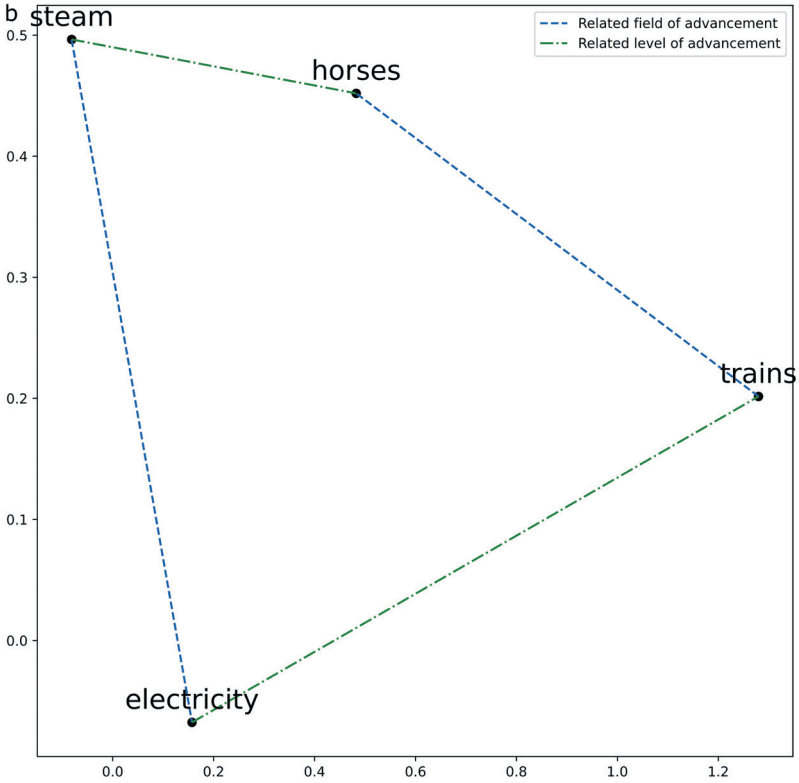


Figure 5.2b: Example of word embeddings as points in space, produced from the word embedding model trained on the 1880s newspaper subcorpus. The coloured lines represent a possible semantic interpretation of the position and relative distance between each point. Credit: Nilo Pedrazzini.

$$v_{england} = \frac{v_{england} + v_{eugland} + v_{ngland} + v_{rngland}}{4}$$

OCR errors can dramatically impact the usability of word embeddings for the linguistic analysis of historical data.[22] Figure 5.4 shows the ten words with the most similar vector (i.e. the ten 'nearest neighbours') to the vector for 'machine' in the 1860s before and after merging the most likely orthographic variants of the word resulting from OCR errors. It is evident that, prior to merging OCR variants, nearly all the ten nearest neighbours of 'machine' are clearly OCR errors of the same word, whereas after OCR post-correction we mainly obtain types of 'machine'.

1860s
england 0.4620602 0.0734637 0.11048217 ... -0.07439072 0.5362225 0.13239789
house -0.02008035 0.2608145 0.06449803 ... 0.24722888 0.6656389 -0.869745
machine -0.08218722 -0.017264733 0.5045748 ... -0.037226044 0.92105055 -0.5772597
train 0.41214764 0.41782025 0.01250272 ... 0.13245606 -0.12823288 -0.45552897
1870s
england 0.48219404 -0.1231414 0.20491965 ... -0.28698 0.6393695 -0.01908828
house -0.36104235 0.31483218 0.41579133 ... 0.48085102 0.68791157 -0.6756131
machine -0.19120322 -0.098553054 -0.31822553 ... 0.616318 1.0226573 -0.54438686
train 0.84905493 -0.008851998 0.0037651716 ... 0.3155296 0.099452145 -0.13350047

Figure 5.3: Examples of vector representations of four words from the diachronic word embedding models trained on the 1860s and 1870s newspaper subcorpora. While the embeddings have 200 dimensions, only the first three and last three dimensions of the vectors are shown. Credit: Nilo Pedrazzini.

However, the fact that OCR errors of a word are represented by very similar word vectors can be seen in itself as indirect evidence that the word vectors are capturing the semantic properties of the words in our corpus, despite the great amount of noise in the data. Spelling variants of a word, if not in themselves indicative of any differences in usage, are expected to have similar vector representations, which are generated using the distribution of the word form in the texts (i.e. their occurrence in the proximity of other particular words) as a reflection of, or proxy for, their meaning. That is, 'maohines' and 'machines' appear in roughly the same location in the model.

Computational exploration

In our exploration of the linguistic change catalysed by the Industrial Revolution, we compared the vector representation of selected words in the most recent decade (the 1910s) with each of the previous time slices.[23] We used a well-established algorithm for detecting the number (if any) of significant changepoints in time-series data, to recognise

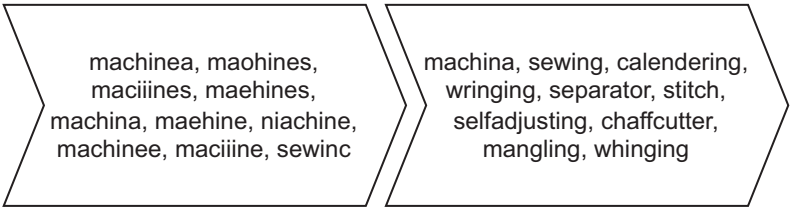


Figure 5.4: The ten-nearest neighbours to 'machine' in the 1860s pre- (left) and post- (right) OCR-error correction. Credit: Nilo Pedrazzini.

potentially major and/or sudden changes in the semantics of the selected words at any point in the nineteenth century.[24]

As shown in Figure 5.5, not only does the vector representation of the words 'train' and 'railway' become progressively more similar to the one in the 1910s (which is generally to be expected) but we also see a much more sudden change after the 1830s for 'train' and after the 1860s for 'railway'. This corresponds to what we know of these words' semantics from external, curated knowledge bases such as the *OED*, which record several new senses around the respective decades. As shown in Figure 5.6, other words, such as 'gear', show a more uneven trajectory, with several peaks and dips (or 'troughs', in time-series terminology), and an overall more gradual, stepwise change, rather than one major semantic-shifting decade.

Overarching comparisons between the paces with which words changed can also be obtained by clustering the semantic trajectories, such as those represented in Figures 5.4 and 5.5, using time-series analysis techniques. This can provide a way of making data-driven decisions about the next steps in the analyses, such as the kinds of variables the linguist, or historian, may want to focus on next. The observations made about the different types of semantic trajectories of 'train', 'railway' and 'gear' (as more abrupt versus more gradual), can be made at scale on several more words by applying a well-known time-series clustering technique called *k*-means with Dynamic Time Warping, which takes the semantic trajectory of words (as cosine similarity scores) as input and clusters them based on this trajectory.[25] In Figure 5.7, Cluster 1 corresponds to a more gradual semantic change trajectory, such as that of 'gear' shown above, whereas Cluster 2 corresponds to a trajectory with at least one major abrupt shift in the semantics of the word, as is the case with 'train' and 'railway'.

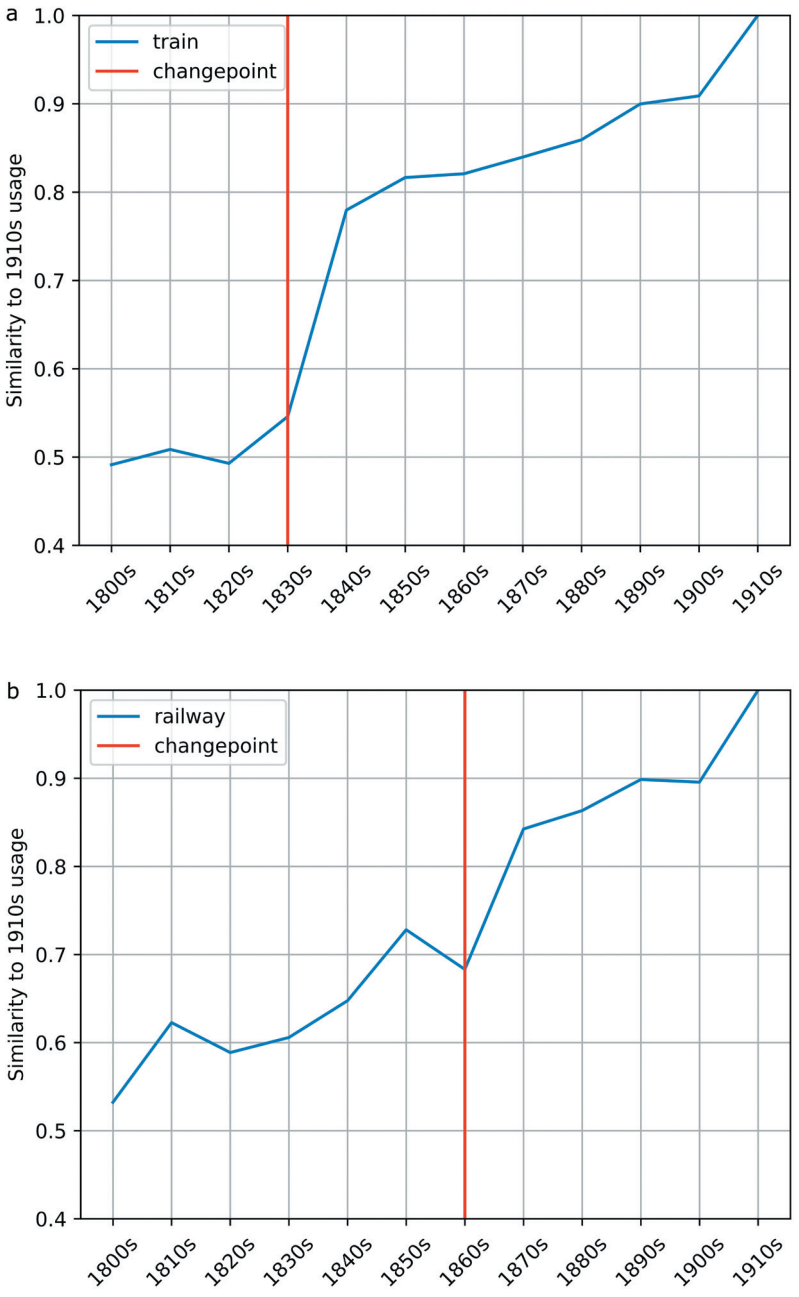


Figure 5.5: Automatically detected changepoints in the semantic change trajectory of a) 'train' and b) 'railway'. Credit: Nilo Pedrazzini.

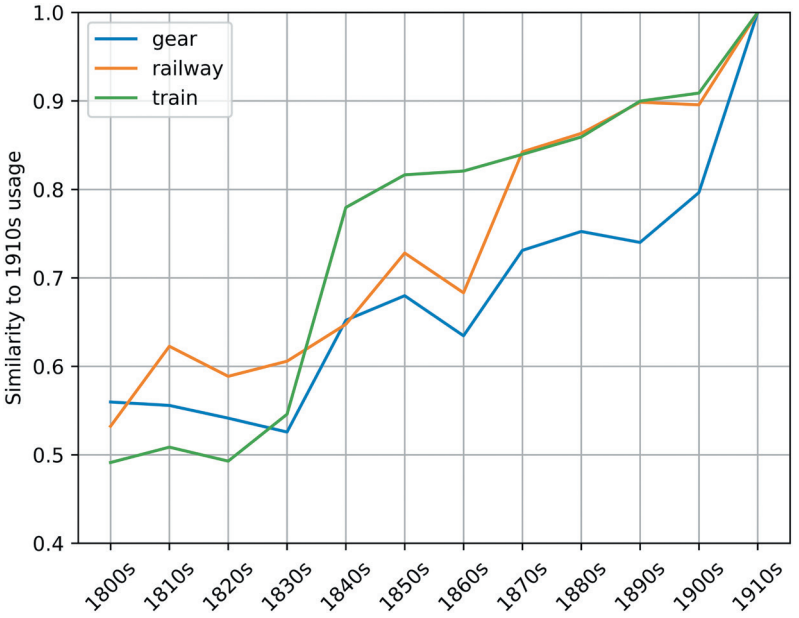


Figure 5.6: Comparison between the semantic change trajectories of three words: 'gear', 'railway' and 'train'. The line represents the similarity between the word vector in each decade and the word vector in the most recent decade (the 1910s). Peaks in the trajectory signify moments in the nineteenth century in which the word acquired meanings found at the end of the period of interest. Credit: Nilo Pedrazzini.

Together with 'gear', in Cluster 1 we find words such as 'bike', 'bulb', 'car', 'match', 'machine', 'stock', 'trolley' and 'trade'; in Cluster 2, next to 'train' and 'railway', we see 'coach', 'stamp', 'traffic' and 'wheel'.

Close-reading interpretation

The automatic changepoint detection method can be complemented with a close-reading interpretation of the nature of the semantic changes by comparing the sets of nearest neighbours of the relevant word before and after any detected changepoint. The changepoints detected for words such as 'coach', 'gear', 'railway', 'train' and 'traffic', for instance, are clearly connected to fast technological changes and their roles in the burgeoning transport system. Most often, the linguistic change manifests itself as an expansion of earlier sets of meanings, showing a transition from pre- to post-industrial usages.

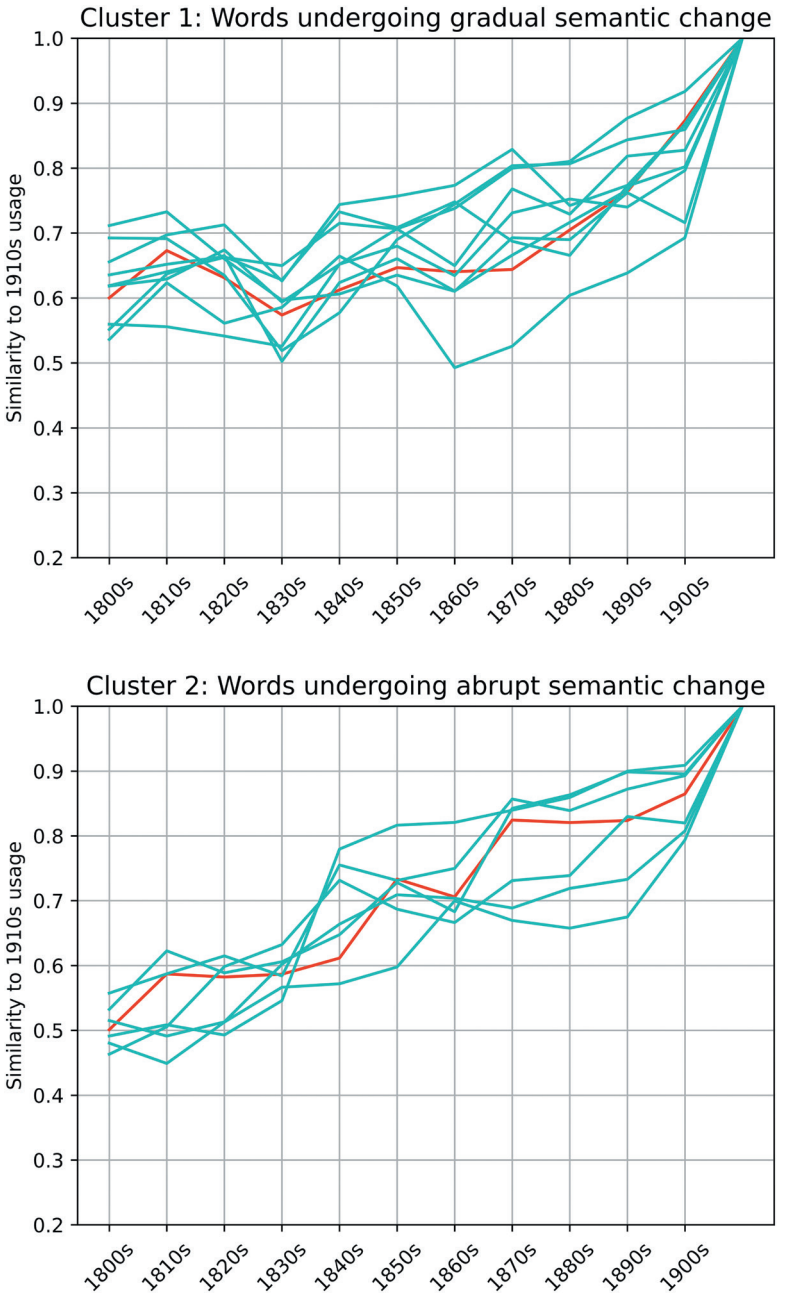


Figure 5.7: Application of a time-series clustering technique (*k*-means with Dynamic Time Warping) to semantic change trajectories. a) the 'more gradual change' cluster of words, including 'bike', 'bulb', 'car', 'gear', 'match', 'machine', 'stock', 'trolley' and 'trade'. b) the 'more abrupt change' cluster of words, including 'coach', 'stamp', 'railway', 'traffic', 'train' and 'wheel'. Credit: Nilo Pedrazzini.

However, not all meaning changes occurred at the same rate, as already mentioned. The semantic journey of 'train' from its initial association with long pieces of fabric, such as robes, skirts or dresses, to its identification with railway transportation showed a rapid, almost complete overtaking by the latter usage of the former in the span of just two decades (see Figure 5.8).

Meanwhile, the word 'railway', whose change is closely associated with that of 'train', showed a somewhat more gradual overtaking by the new meaning, which remained as widespread in newspapers as the earlier usage (more generically as 'any roadway along which wheels may run to facilitate the transport of heavy loads') for at least three more decades after the sense specifically associated with trains became predominant around the 1860s (Figure 5.9).

We also captured much more gradual and complex changes. The trajectory of 'gear' (Figure 5.10) shows an incremental introduction of specific senses related to new mechanical advances throughout the nineteenth century. This, in turn, indicates that its semantic change has to

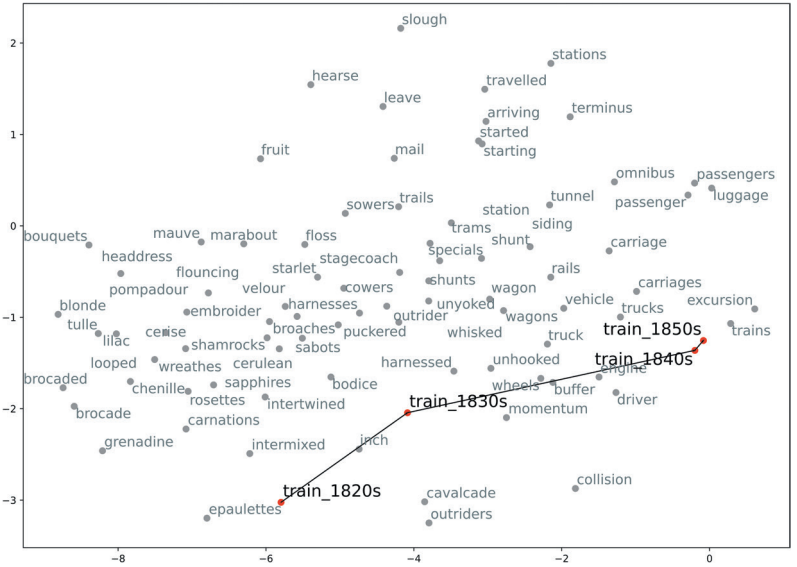


Figure 5.8: Semantic change trajectory of 'train' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century. Credit: Nilo Pedrazzini.

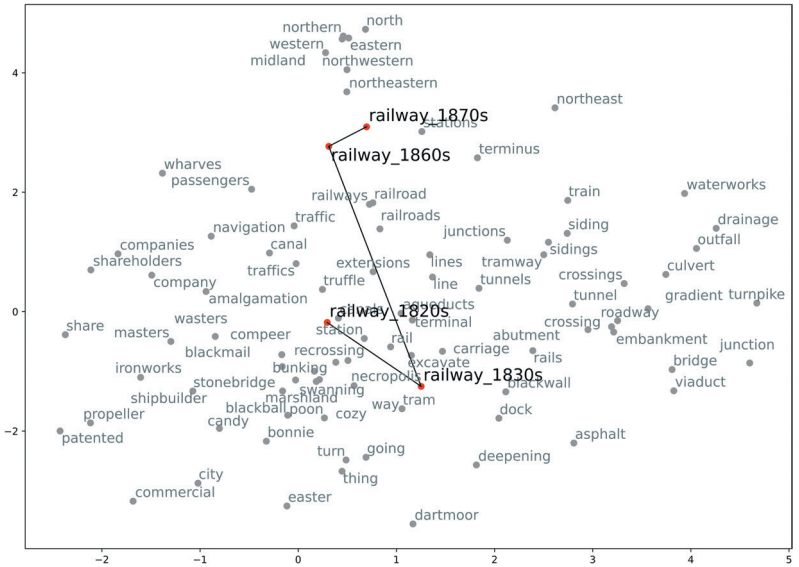


Figure 5.9: Semantic change trajectory of 'railway' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century. Credit: Nilo Pedrazzini.

do with different associations with other words becoming predominant at various points in time (e.g. 'ironing gear', 'caulking gear', 'winnowing gear', etc.).

This method also allowed us to highlight potential discrepancies with existing linguistic resources, such as the *OED* and its thesaurus, in the nature and timing of semantic changes. For instance, the evolution of 'coach' (Figure 5.11), which emerged from automatically detecting change-points, challenged the *OED*'s classification of its nineteenth-century use as a 'railway carriage' as predominantly American English. Its semantic shift appears to be attested in England as well and may have occurred earlier than previously documented, offering a potential refinement to historical linguistic narratives. As we discuss in the 'Visualisations and Analysis' section, these shifts are entangled with broader linguistic and cultural processes, such as the influence of American English, and the international press system's role in meaning diffusion, exposing the irreducible, non-linear complexity of language change.

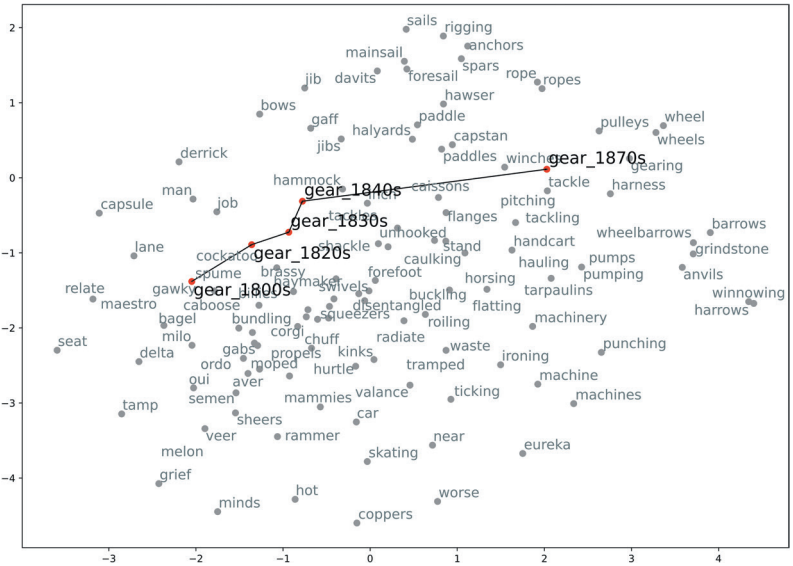


Figure 5.10: Semantic change trajectory of 'gear' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century. Credit: Nilo Pedrazzini.

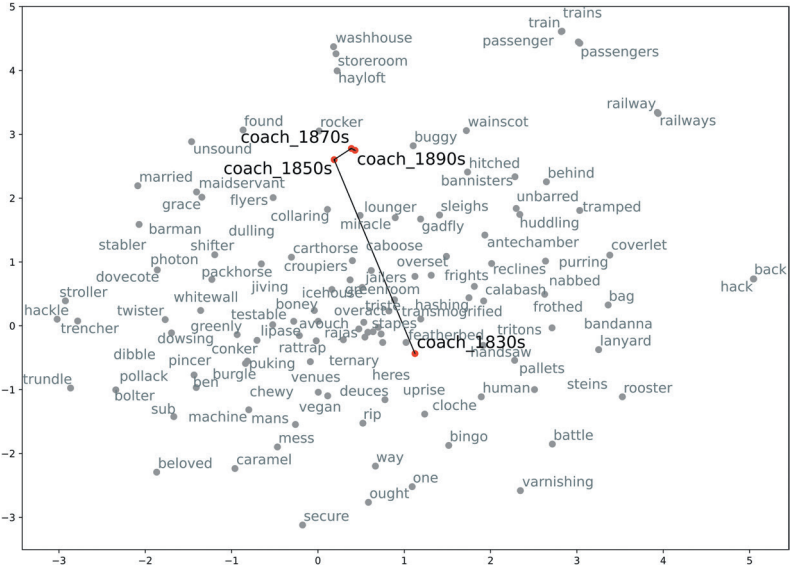


Figure 5.11: Semantic change trajectory of 'coach' as reflected in the changing set of nearest neighbours (most similar word vectors) at various points in the nineteenth century. Credit: Nilo Pedrazzini.

The potential of newspaper metadata: preliminary insights on regional semantic shifts

Our analysis is only one possible way of exploring the multi-billion-word newspaper corpus. In particular, we have only begun to uncover the full potential offered by the extensive metadata that has been linked to each newspaper title, thanks to Living with Machines' digitisation of nineteenth-century press directories and the complex process of data extraction and linkage undertaken as part of the Environmental Scan, outlined in Chapter 1.[26] This enhanced metadata for the newspapers includes diverse variables such as the geographical region, the declared political orientation of the newspapers, and their price at different points in time, all of which can provide useful insights into the complex interrelations between semantic change, technological innovation and language use.[27]

In a preliminary experiment that illustrates the potential of this enhanced corpus, we used the place and year of publication of each newspaper to divide the corpus into two subcorpora containing articles published in two broad geographical regions, North and South England, historically corresponding to one of the main socio-political divides in Britain, before further splitting each geographical subcorpus into temporal subcorpora. After automatically detecting potential changepoints for words such as 'bulb', 'coach', 'gear', 'matches', 'railway', 'stamp', 'trade', 'traffic', 'train' and 'wheel', we compared the results from the two subcorpora against each other to assess whether their semantic shift occurred virtually simultaneously across the two regions or whether some degree of diatopic (i.e. spatial) variation could be posited. As an exploratory step, we also considered potential differences in usage between singular and plural forms of the same words.

Our preliminary results showed, in fact, differences between the North and South of England. While for some words, like 'match' and 'stamp', a changepoint was detected for both regions, for other words, such as 'machines' and 'stock', a changepoint was only detected for the South, and for others, like 'trade', 'bulbs' and 'cars', only for the North. Moreover, even for words with a potential changepoint in both regions, the decade in which the shift occurred may differ: for 'match', for instance, a changepoint was detected later in the century for the North than for the South. 'Stamp' appears to have had a sudden change in usage in North England from

its older generic meaning of ‘marking’, ‘engraving’, or ‘impression’ to its newer philatelic sense, as becomes clear after inspecting which words are most closely associated with stamp before and after the detected changepoint. Similarly the word ‘cars’ seems to have had a sudden change in the North from its older existing meaning of a wheeled, usually horse-drawn conveyance to that associated with railway carriages or waggons. In the South, the newer meanings of both words are attested but do not induce a changepoint detectable by the algorithm, suggesting a less abrupt and less intense change in usage. Figures 5.12 and 5.13 compare the change in the nearest neighbours of the word ‘cars’ in the North and South of England, respectively.

For some words, a changepoint was detected in a region either only in the singular or the plural form, as in ‘match’ (but not ‘matches’) in the South or ‘stamp’ (but not ‘stamps’) in the North of England; in other cases, a changepoint was detected earlier for one of the two forms. These subtle differences, which warrant further exploration, may have to do with the different usages

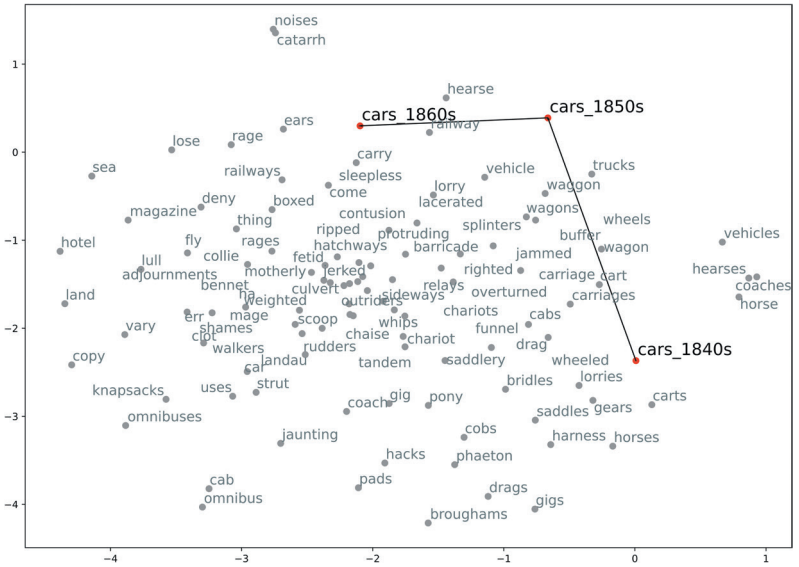


Figure 5.12: Semantic trajectory of ‘cars’ in the North of England, based on the nearest neighbours of this word in the respective geographic subcorpus in subsequent decades. Credit: Nilo Pedrazzini.

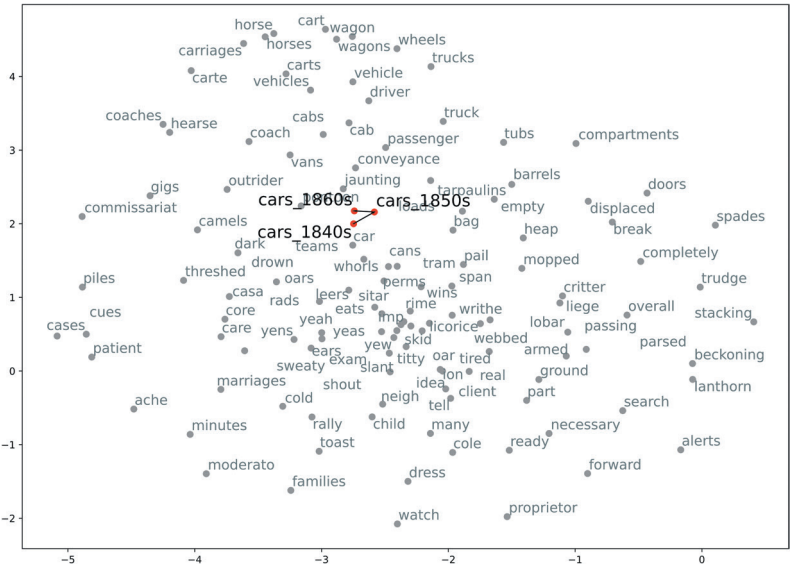


Figure 5.13: Semantic trajectory of 'cars' in the South of England, based on the nearest neighbours of this word in the respective geographic subcorpus in subsequent decades. Credit: Nilo Pedrazzini.

and therefore different triggers for semantic change associated with some words in the singular and plural, for example when they are used as plural generics (e.g. 'cars are an amazing invention'), as opposed to referring to multiple instances of a concrete object (e.g. 'there are cars on the street').[28]

These are, however, preliminary hypotheses and no systematic attempt was made to validate and interpret the results. Only further more rigorous analyses, such as those implemented by Pedrazzini and McGillivray and outlined above, will allow us to establish what factors most strongly drove diatopic variation in the process of semantic change.[29]

Designing sense disambiguation tasks for crowdsourcing volunteers

How did the words 'trolley', 'car', 'coach' and 'cycle' change over time? In this section we describe how we built novel crowdsourcing tasks with the result that hundreds of members of the public helped us find out.

The early results of the changepoint detection work described above fed into our work engaging the public with computational linguistic work via crowdsourcing tasks. In this section, we discuss how we built on other crowdsourced sense disambiguation tasks on the Zooniverse citizen science platform created earlier in the project to design semantic annotation tasks.[30] These tasks invited the public to contribute to our research by annotating individual instances of specific target words with historical senses from the *OED*.

Unlike commercial, Mechanical Turk-style projects with financial rewards, volunteer participants in crowdsourcing and citizen science are motivated by taking part in an enjoyable and meaningful task with interesting collections.[31] It is challenging to design tasks that both make a contribution to a wider issue or research area and are clear and interesting enough to motivate ongoing contributions. Our work on the 'Language of Mechanisation' workflows built on the lessons learnt from earlier collaboration on 'What was a machine?' workflows, which included our first semantic annotation task and a text summarisation/transcription task.[32] It also provided an opportunity to explore the potential of voluntary crowdsourcing to provide data for Digital Humanities research such as the development of computational linguistics algorithms.

The changepoint detection analysis described above provided a list of potential target keywords of interest. We reviewed the historical senses listed for each entry in the *OED*, looking for senses that were distinct enough to make it easy for volunteers to choose the right sense for a given word in context. While we considered a number of words, ultimately we designed tasks around the words trolley, car, coach and cycle (bicycle or motorcycle). We set up four 'workflows' (a group of sequential tasks on a specific image from a historical newspaper), each focused on a specific target keyword. See Figure 5.14 for an example.

Having established relationships with volunteers through earlier projects on Zooniverse, we were able to publicly share our work in progress and invite feedback from them.[33] Their comments helped refine our design and explanation of novel tasks that required volunteers to read historical texts closely. We undertook a range of outreach and marketing activities when the workflows were ready to officially 'go live'.[34] Volunteer contributions via our workflows created a rich set of annotations, which afforded the analysis described below.

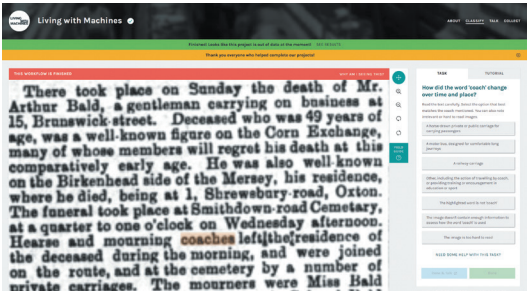


Figure 5.14: The Zooniverse task interface showing newspaper articles and the related annotation options for 'coach'. Credit: Mia Ridge; British Library.

The annotated articles are also available for download and re-use in the British Library's research repository and have been extensively documented in readme files and a data paper.[35] For further details on the results see Appendix 1.

Visualisations and analysis

Following the crowdsourcing work described above, we shared the data with King's Digital Lab (KDL, King's College London) so that they could help analyse and visualise the corpus. The KDL team collaborated with the Living with Machines project team in the late summer of 2023 over roughly two months to produce three products in the form of Notebooks which were used to inform the analysis that follows. The design and implementation of the Notebooks is detailed in Appendix 2.

The evolution of 'trolley': from hand-powered cart to fuel-powered waggon

According to the *OED*, early examples of the word 'trolley' in English usage represent local, vernacular words for a low, narrow cart, particularly one designed for use in constricted spaces.[36] The *OED* editors cite Edward Moor's book *Suffolk Words and Phrases* (1823) as the first attested usage, underscoring the word's regional dimension in the early nineteenth century. Our analysis appeared to throw up three examples before this date, but manual inspection suggests that these are all false positives. Our first confirmed uses come from 1850, when we have two cases. In both cases the local newspaper felt the need to explain the word's

meaning, suggesting that 'trolley' was still viewed as a neologism at this point (both instances explained it was a type of waggon, one specifying what type: 'in the trolley, or low waggon'). But, as the *OED* records, by the 1850s 'trolley' (or sometimes 'trolly' or 'trawley') was also being used more specifically to refer to low waggons designed to run on rails or other types of track, either on the railways or in and around factories and docks. Interestingly, the present-day dominant meaning, where 'trolley' is used to denote a hand-powered means of conveyance on small wheels or castors (as in tea-trolley or supermarket-trolley), arose much later, beyond the period covered in our analysis (the *OED*'s first attested use of this sense is an account of someone purchasing a tea-trolley in 1937). Although both early senses exist in parallel from the 1850s, our analysis shows that the balance of usage shifted across the period 1850–1920.

In our analysis we therefore focus on the shifting balance between uses of 'trolley' that refer to a small waggon pulled by human or animal power, and those referring to a steam or electric-powered vehicle, usually mounted on rails.

As Figure 5.15 extracted from the project's Observable Notebook shows, the number of annotations available for the early years of our corpus (before 1860) is very low, and this holds across the board for all annotation categories identified, that is, the pre-industrial sense of 'trolley', the mechanisation sense, and other, as well as the cases in which the annotators were not able to make a decision. The increase in annotations after 1860 corresponds to the larger size of the more recent portion of our corpus, reflecting the growth of the press and its evolving approach to vernacular usage.

When we consider the relative frequencies of the two main senses considered we can assess their complementary distribution in the corpus. By factoring out the growing size of our corpus we can focus on the trajectory of the pre-mechanisation and post-mechanisation meanings.

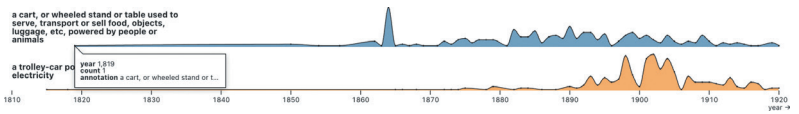


Figure 5.15: Raw total count of annotations for each word meaning ('trolley' as cart vs. trolley-car) in each year. Credit: King's Digital Lab.

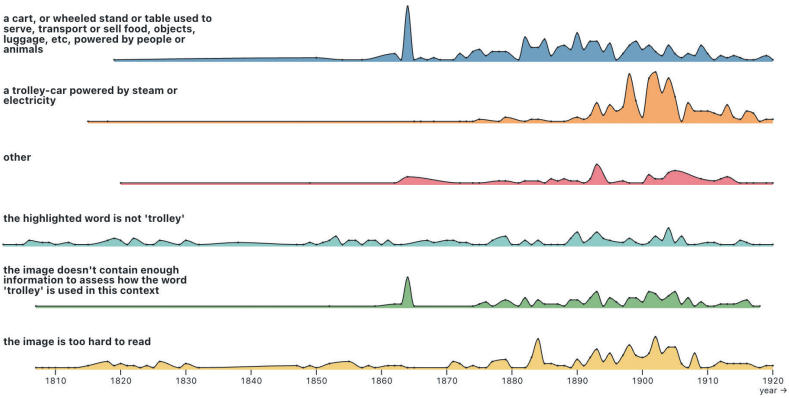


Figure 5.16: Raw total count of annotations for each word meaning ('trolley' as cart vs. trolley-car vs. other) and other classifications in each year. Credit: King's Digital Lab.

We excluded the 'Other' label from this analysis as this category grouped together annotations which displayed a level of ambiguity in the contexts. Figure 5.16 shows the number of annotations for each meaning divided by the total number of annotations available for 'trolley' in each year. It clearly shows how the sense 'a trolley-car powered by steam or electricity' overtook the original sense in the late nineteenth century.

In Figure 5.17 the curves in the plots are created using a monotone-x curve, that is, a curve interpolation method that ensures a smooth increasing or decreasing curve when visualising data. The trend lines (straight) in the plots represent the general direction or pattern of the annotation counts for that word per year and are generated using linear regression. The accuracy of these trend lines is affected by the limited size of the dataset before 1860 and therefore should be interpreted with caution. Nevertheless the main trends confirm our expectations, with the sense related to mechanisation gradually taking over the original one.

The shaded area in the plot represents the potential variability or imprecision in the data, which explains why they show values above 100 per cent or below zero per cent. The range is calculated using confidence intervals for proportions.

The meaning of 'trolley' as cart powered by people or animals is mainly associated with a North American use; however, the annotations data

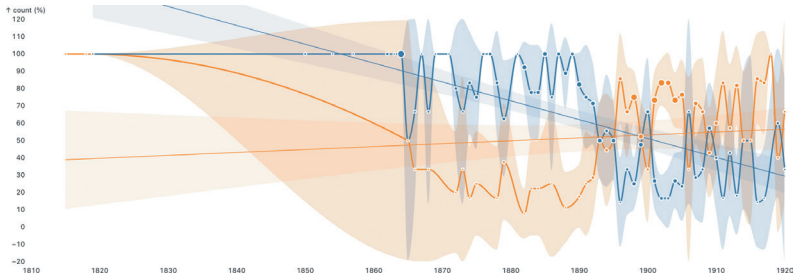


Figure 5.17: Relative frequencies of the annotations for each of the word meanings for the word 'trolley' (blue for trolley as cart and orange for trolley as car). Credit: King's Digital Lab.

under examination seem to show this usage to be dominant in British sources too. British local newspapers varied considerably in the extent to which they reported national and international news, but it seems doubtful whether they gave extensive coverage to American stories involving non-mechanical trollies. More likely this reflects regional variations in vernacular usage of the word.

Given the relatively small number of confirmed uses we are dealing with for the word 'trolley' (561 articles), any overall conclusions must be tentative. In Figure 5.18 our statistical analysis identifies two key moments of semantic change, 1860 and 1890. The change point detection algorithm is a method used to detect abrupt changes or shifts in time-series data. It is useful when dealing with data that exhibits constant behaviour, where the data remains relatively constant between certain points in time but undergoes sudden changes at other points. The first point, in 1860, seems to be associated with the widespread adoption of the regional vernacular word 'trolley' as a word to describe a low waggon used for moving people or goods. The second, in 1890, appears to be associated with the shift towards meanings associated with powered vehicles, even though Britain did not adopt the American term 'trolley-car', preferring to call its powered public transit vehicles 'trams'. Manual inspection suggests two reasons for this. First, British newspapers carried extensive reports of American news which tended to retain the vernacular usage, perhaps for added colour or authenticity. Second, specific parts of the tram, notably the connecting rod that picked up electric current, adopted the American usage; we find frequent reports about incidents involving the 'trolley-pole' of British municipal

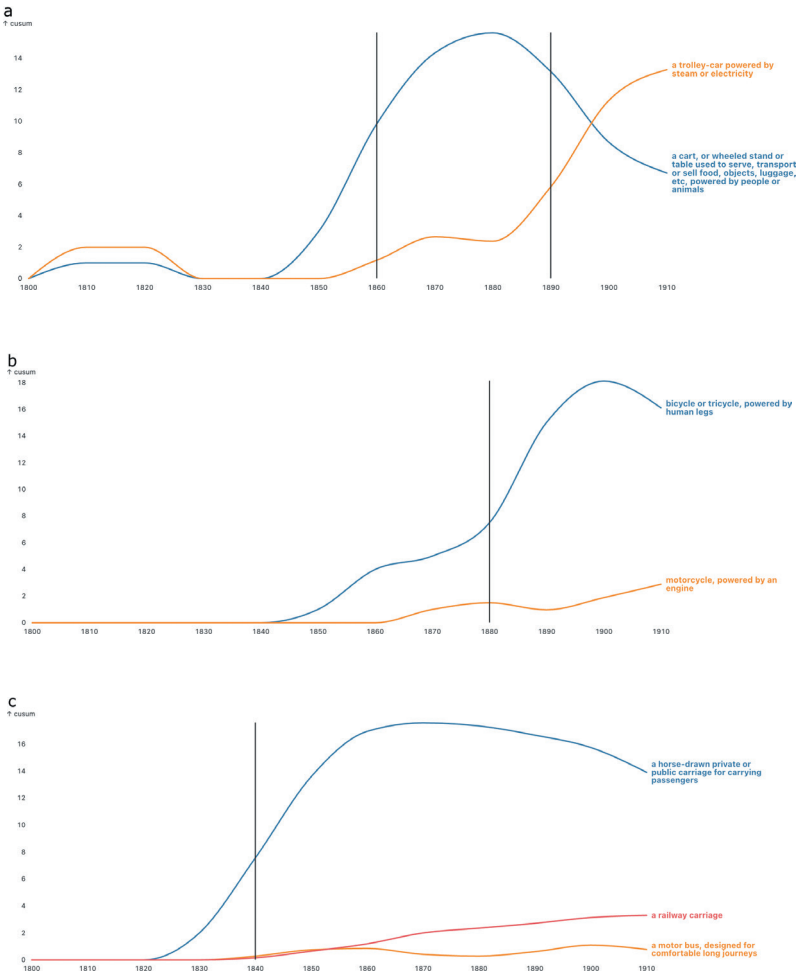


Figure 5.18: Decades when the words a) 'trolley', b) 'bike' and c) 'coach' may have changed meaning according to our changepoint detection analysis. Credit: King's Digital Lab.

trams. Together, these two influences, the allure of American news and the adoption of the American term for parts of British trams appear to explain the meaning shift about 1890.

Mechanisation and meaning shift in other transport words: 'bike', 'coach' and 'car'

We repeated this analysis with three additional words connected with transport that are known to have undergone significant semantic change, and where mechanisation was thought likely to have played its part in these linguistic processes: 'bike', 'coach' and 'car'. Of these three, only 'bike' was a neologism, being an Anglo-Saxon shortening of 'bicycle', which the *OED* records was itself borrowed from French in the 1860s at the height of the first cycling craze (the first English use of 'bicycle' cited by the *OED* is a *Daily News* story from 1868).[37] The earliest recorded use of the shortening 'bike' is from a cycling trade magazine in 1880, a few years before the great *fin de siècle* cycling boom generated by the invention first of the safety bicycle and then the pneumatic tyre in the 1880s (see Chapter 4).[38] The *OED*'s first attested usage of 'bike' to mean motorcycle is from *Automobile Topics*, a specialist American motoring magazine, in 1903; their first British example comes from a letter written by T.E. Lawrence to George Bernard Shaw in 1924. Once again mechanisation and Americanisation appear to be deeply intertwined in this process of semantic change. It seems likely that, in the era before Hollywood and Tin Pan Alley exerted global influence, newspapers and magazines may have represented important vehicles driving semantic change across the Anglophone world. Important work has been done tracing how news stories spread (and evolved) across continental and global print and telegraph networks; our focus on semantic change suggests ways to trace the effects of these radical new connections linguistically.[39]

The other two transport words we analysed, 'coach' and 'car', both have deep roots in British English as nouns describing vehicles for conveying people and/or goods. The *OED*'s earliest examples of 'coach' meaning a large, horse-drawn carriage date from Elizabethan times (as 'coche' and 'cosche'); their first usage with the spelling 'coach' is from the diary of Samuel Pepys in 1665.[40] The origins of the word 'car' are even older, with the first attested uses (with various spellings) dating from the fourteenth and fifteenth centuries. Like 'coach', it was a word for a horse-drawn conveyance and could be used to refer to anything from a humble cart to a royal carriage.[41] Both words registered significant semantic change in the age of mechanisation, with North American interconnections again playing a part in this story. In British English, 'coach' was initially adopted to denote the

covered, especially first-class, railway carriages on early trains (at a time when poorer passengers were forced to travel in open waggons exposed to the elements and the engine's soot and smoke). Later in the nineteenth century, it became used more generally to denote any railway carriage (all passengers now enjoyed cover), but phrases like 'Pullman coach' retained the earlier sense of luxury. The predominant present-day meaning (i.e. a comfortably equipped motor bus designed for long journeys) appears to have had its origins in pre-1914 America; the first British usage cited by the *OED* dates from 1930.[42] In British English, the *OED* records that the meaning of 'car' was extended to denote the passenger compartment of various new modes of transport from the late eighteenth century, including air balloons and cable cars, but unlike in America, the word was not widely used to denote railway passenger compartments (instead 'coach' or 'carriage' became the preferred terms). The *OED*'s earliest citations for the dominant present-day usage of 'car', both in Britain and North America (i.e. a motorised, private road vehicle for people), date from 1896 and 1900, when the motor car first began to grip the public imagination in industrialised societies.[43] Crucially, it appears that all three words continued to be used interchangeably to denote both mechanised and non-mechanised vehicles, with lexical context the only way to signal the intended meaning. Our aim was to use diachronic analysis to map how the balance of semantic meaning shifted in local newspapers across the long nineteenth century.

Some of the Notebook graphs are helpful to offer insights and spur reflection on the decades when changes of meanings occur (see Figures 5.19 and 5.20). For example, in the case of 'bike', the annotations data seem to support a claim that its meaning as motorbike, powered by an engine, occurs quite late around 1910, while the meaning of bike powered by human legs seems to take off around 1880. This predates the technical innovations that fuelled the great cycle mania of the 1890s, when the bicycle finally became a means of popular mass transport, reminding us that linguistic culture can shift ahead of the culture of everyday practice.[44] Since motorbikes became commercially available from the mid-1890s one might also infer that linguistic culture could also prove resistant to change once a new usage became embedded in everyday practice. It was during the interwar period that motorbikes (briefly) became the predominant means of private motorised road transport in Britain, creating the context for the newer meaning of 'bike' to take hold.[45]

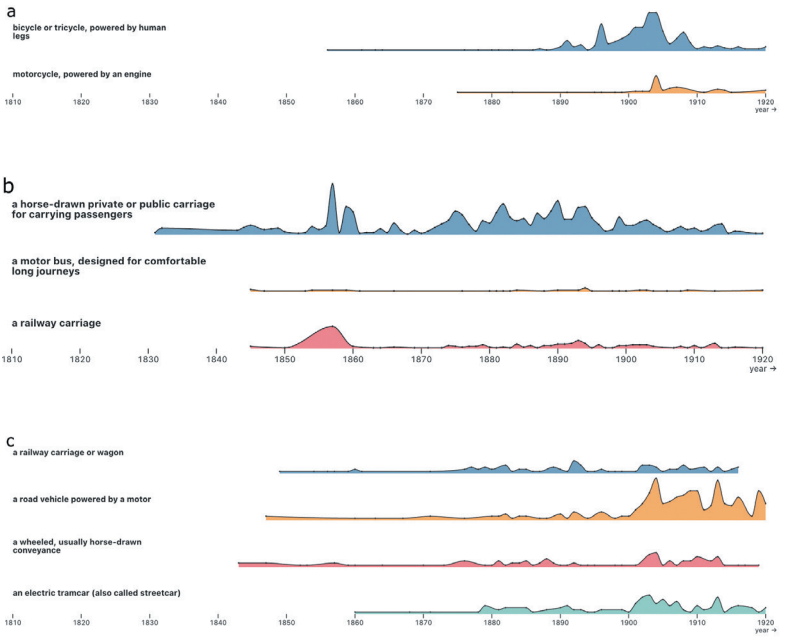


Figure 5.19: Raw total count of annotations for each word meaning associated to a) 'bike' (powered by human legs vs. an engine), b) 'coach' (horse-drawn carriage vs. motor bus vs. railway carriage) and c) 'car' (railway vs. road vehicle powered by a motor vs. wheeled horse-drawn conveyance vs. electric tramcar) respectively, in each year. Credit: Barbara McGillivray and King's Digital Lab.

According to the annotations data at hand, new meanings for the word 'coach' seem to occur in the decade starting in 1840; however, these clearly do not take over the horse-drawn carriage meaning; indeed, other words for railway 'coach' and motor 'coach' were in use, notably 'carriage' and 'charabanc' respectively, leaving the older meaning of 'coach' dominant throughout this period (see Figure 5.21). An onomasiological analysis which looks at the various words used to express a given concept would be needed to substantiate these insights. In Britain, the motor bus meaning of 'coach' would probably be found to come into dominant usage when long journeys became possible, in the interwar period.

With respect to 'car', not surprisingly, the road vehicle meaning started to take off in the early 1900s (see Figure 5.22). In Figure 5.23 the parameters have been tweaked to detect further change points.

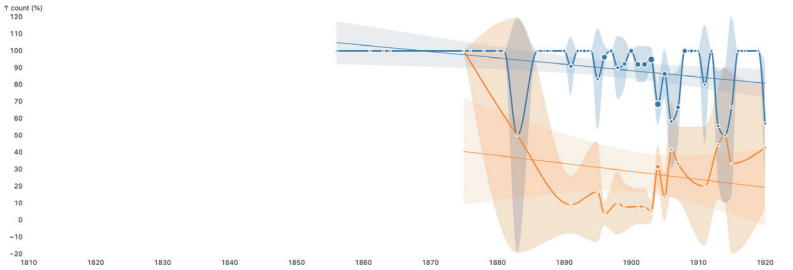


Figure 5.20a: Relative frequencies of the annotations for each of the word meanings for 'bike' in each year. The blue line represents 'bicycle or tricycle, powered by human legs'. The orange line represents 'motorcycle, powered by an engine'. Credit: Barbara McGillivray and King's Digital Lab.

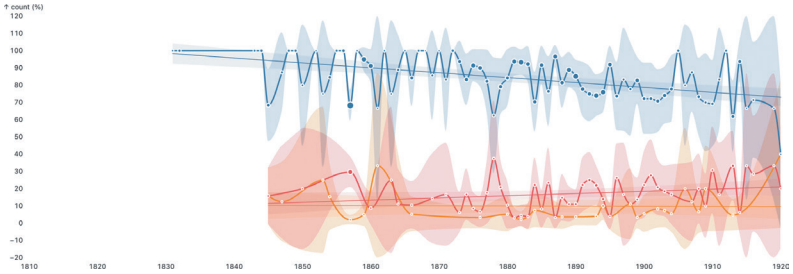


Figure 5.20b: Relative frequencies of the annotations for each of the word meanings for 'coach' in each year. The blue line represents 'a horse-drawn private or public carriage for carrying passengers'. The orange line represents 'a motor bus, designed for comfortable long journeys'. The red line represents 'a railway carriage'. Credit: Barbara McGillivray and King's Digital Lab.

A first change point occurs around 1850 where the two new meanings (electric tramcar and road vehicle, not yet the private motor car) appear and begin to rise; these remain subdued, however, until the second change point around the 1900s, where the motorised road vehicle meaning takes off and other meanings diminish or remain at a low level. This change point marks the explosive growth of the private motor car as a presence not just on British roads, but in the national psyche. This is the moment that saw Kenneth Grahame create Mr Toad as a demonic, car-obsessed road hog in *The Wind and the Willows* (1908), and road safety concerns fed the call for drivers' licences, speed limits and measures to control the speeding driver.[46]

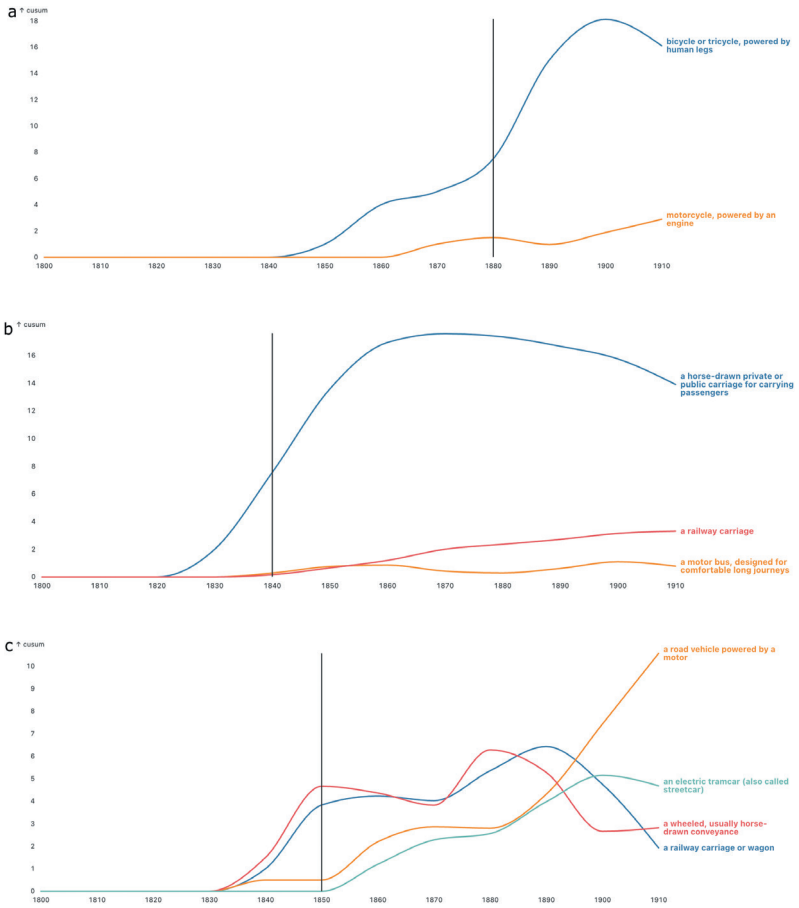


Figure 5.21: Decades when the words a) 'bike', b) 'coach' and c) 'car' respectively may have changed meaning according to our changepoint detection analysis (window set at 2 and penalty at 25). Credit: Barbara McGillivray and King's Digital Lab.

Conclusion

In this chapter we reported on an exploration of semantic change during the era of mechanisation. Semantic change is a highly complex phenomenon, where external factors, such as cultural influences and the adoption of terminology from different regions, significantly impact the semantic trajectory of words. At the same time, words do not change meaning in isolation of each other, and connected changes add to this complexity.

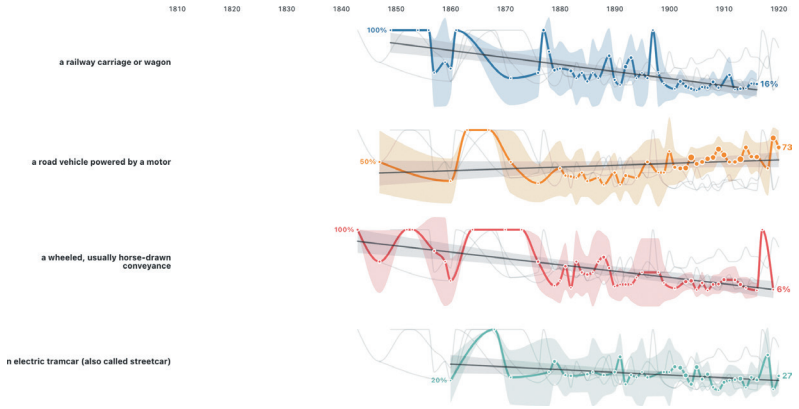


Figure 5.22: Relative frequencies of the annotations for each of the word meanings for ‘car’ in each year. Credit: Barbara McGillivray and King’s Digital Lab.

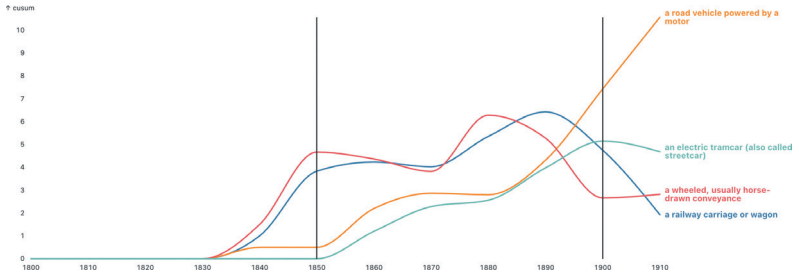


Figure 5.23: Potential decades when the word ‘car’ changed meaning with two change points represented by the vertical lines (penalty decreased to 22). Credit: King’s Digital Lab.

Our analysis leveraged computational algorithms, manually annotated data and close reading to add a new quantitative dimension to our notion of semantic change in the mechanisation era. This work highlights how computational tools, when combined with historical analysis and manual annotation, can significantly enrich our understanding of language change, providing a new level of nuance compared with traditional methods alone, and uncovering hidden shifts, including those linked to regional variation.

Our computational approach, based on diachronic word embeddings, enabled us to identify both gradual and abrupt semantic shifts in key mechanisation-related terms over the nineteenth century such as ‘train’,

'railway' and 'gear', shedding light on the nuanced trajectories of linguistic change catalysed by industrialisation. The employment of time-series analysis techniques allowed us to detect significant changepoints, offering insights into the temporal dynamics of semantic evolution. Furthermore, close-reading interpretation complemented automated methods, providing rich contextual understanding of the semantic shifts. These insights partially align with historical accounts of the change in the English lexicon in response to advancements in transportation and machinery, reinforcing the notion of language as a dynamic reflection of socio-cultural and technological change. For example, the semantic trajectory of 'train' from its earlier association with fabric to its current usage in railway transportation mirrored historical narratives of the transformative impact of steam locomotion on transportation systems. Similarly, the semantic evolution of 'gear' demonstrated incremental changes in meaning corresponding to advancements in mechanical engineering and industrial production.

At the same time, our analysis revealed semantic shifts not extensively documented in existing linguistic resources and dictionaries. This is particularly evident in the case of 'coach', where we showed that its nineteenth-century usage as a 'railway carriage', attributed predominantly to American English in the *OED*, had significance in English discourse as well, potentially occurring earlier than previously recognised. Our investigation extended beyond a diachronic analysis to incorporate regional variations and explore diatopic shifts in meaning, by leveraging foundational work undertaken on the Living with Machines project as part of our Environmental Scan (see Chapter 1). Our preliminary findings identified potential differences in the semantic trajectories between North and South England, pointing to the interplay between geography, language use and technological innovation during the mechanisation era, and stressing the importance of considering socio-geographic factors in the study of linguistic change across different linguistic communities.

The voluntary crowdsourcing annotation process leveraged the collective efforts of multiple annotators to annotate individual instances of *trolley*, *coach*, *car* and *bike* in nineteenth-century British newspapers with historical senses from the *OED*. The design of these tasks involved collaboration between researchers and volunteers to develop the first experiment in crowdsourced semantic annotation of historical texts that we are aware of. The crowdsourced annotation data provided the starting point for a series of interactive visualisations that allowed us to examine words

potentially changing meaning during the nineteenth century. The process involved several iterations, integrating user research, project requirements and technical solutions. The visualisations found evidence of the semantic shifts that the words of interest underwent and highlighted some of the complexities inherent in any empirical analysis of semantic change. At the same time, the low number of annotated instances, particularly in the early years of our corpus, calls for some caution in the interpretation of the data, while the growing size of our corpus in the second half of the nineteenth century reflects the broader expansion of the press during that period and possibly its opening up to vernacular usage. This combination of computational algorithms and crowdsourced annotation sets a new precedent for how we study historical texts, opening up new possibilities for tracking semantic shifts in vast digital corpora. Thanks to the development of interactive visualisations and the use of diachronic word embeddings, we have paved the way for more dynamic, quantitative approaches to corpus linguistics, moving beyond traditional dictionary-based methods to explore the evolving nature of language in its social context.

The analysis of 'trolley' found evidence of the newer meaning 'a trolley-car powered by steam of electricity' overtaking the earlier meaning in the late nineteenth century, with two key moments of semantic change identified around 1860 and 1890, corresponding with shifts in regional vernacular usage and influences from American English. A similar connection between semantic change and Americanisation affected the neologism 'bike', which originally referred to human-powered bicycles but later extended to encompass motorbikes, particularly during the interwar period when motorbikes gained popularity. On the other hand, 'coach' had its origins in large, horse-drawn carriages from Elizabethan times, and over time was used to refer to covered railway carriages on early trains and eventually evolved to denote motor buses for long journeys. Similarly, 'car' initially described various horse-drawn conveyances, later evolving to include passenger compartments of new transport modes. The shifts in meaning of the three terms 'bike', 'coach' and 'car' correlated with the rise of motorised road vehicles and the popularisation of the private motor car, although this transition was not linear or uniform, as all were employed interchangeably to refer to both motorised and non-motorised vehicles, relying on contextual clues to convey the intended sense. For these reasons, our diachronic analysis aimed to chart the evolution of semantic nuances in the newspapers throughout the extended nineteenth century.

While our analysis has provided valuable insights into the semantic evolution of mechanisation-related vocabulary during the nineteenth century and has helped paint a more nuanced picture than a mere binary shift from an older meaning to a newer one, it is essential to acknowledge the limitations of our approach. Despite efforts to mitigate noise, the presence of OCR errors in historical texts posed a significant challenge, potentially impacting the accuracy and reliability of our semantic analyses. Although we implemented post-correction techniques to address OCR inaccuracies automatically, the extent to which these corrections influence the semantic trajectories of words remains a point of uncertainty, underscoring the need for further methodological refinement. Moreover, our preliminary exploration of regional semantic shifts highlighted the complexities of diatopic variation but did not systematically validate or interpret the results. As such, a more rigorous interrogation of regional differences in language use, informed by historical, sociolinguistic and geographical insights, would be needed to contextualise our findings accurately.

Manual annotation and close-reading analysis has highlighted the importance of considering the context, as words were often used interchangeably to denote both mechanised and non-mechanised vehicles, requiring careful interpretation based on lexical context. Given the considerable time resources needed to conduct manual semantic annotation and to ensure high agreement between annotators, its limited extent posed a notable constraint on our results. The limited amount of manual annotation (in part the result of the decision to use open access content) restricted the robustness of the insights we were able to derive from it. Expanding the scope of manual annotation in the future would enhance the richness and accuracy of the results. Additionally, a more extensive manual annotation process would facilitate the validation and refinement of automated algorithms, underscoring the importance of investing in more extensive manual semantic annotation efforts to augment the depth and rigour of semantic change research.

As pointed out in our methodological and process-related discussion, several further valuable lessons have emerged, guiding future work at the intersection between semantic change and historical research. In line with the overarching ethos of the broader Living with Machines project, we stress the significance of interdisciplinary collaboration, without which this work package would have not succeeded. The synergy among computational linguists, historians, research software engineers, library

curators and volunteers has been instrumental to this work, and made it possible to integrate diverse methodological approaches, and manual and automated processes. Our contributions hold a significant potential not only for advancing our understanding of historical linguistics but also for offering a framework for the study of semantic change in historical texts. The result has been a richer and more multifaceted understanding of the evolution of historical language and, with it, of complex historical phenomena.

Appendix 1 – Crowdsourced task results

There were 528 individual volunteers who contributed to the 'Language of Mechanisation' Zooniverse workflows to complete a series of tasks related to the analysis of historical newspaper articles (subjects in the data model below, Figure 5.31). The tasks focused on the analysis of the words 'trolley', 'car', 'coach' and 'bicycle'/'motorcycle' and involved classifying each article with the provided definition (or sense) that best matched the use of that word in context. Data about the newspaper articles came from the digitised newspapers, combined with the information, such as price and political leaning, available in Mitchell's Press Directories and processed by the Living with Machines project. The 'Explore the Data Sources' section of the Observable Notebook around the language of mechanisation presents some of this data as showcased in Figures 5.24, 5.25 and 5.26.

Whenever using human annotated data for research, measuring the inter-annotator agreement, that is, the extent to which the annotators produced the same annotation for the same initial data, helps the researchers assess the reliability and challenges of the annotation.

We calculated agreement rate and frequency as follows. We grouped all annotations available per article. For each article, we calculated the total number of annotations for that article and identified the annotation that was used the most times for the article; we refer to this as 'max'. We calculated the agreement rate by dividing the max by the total of annotations. Finally, we counted the number of times each agreement rate occurs. The Notebook allows the users to set a value for the minimum agreement required. We set this value to 65 per cent to ensure that at least two thirds of the annotators (typically two out of three) agreed. The analysis presented in the 'Mechanisation and Meaning-Shift in Other Transport Words' section is based on the annotations for which this minimum value was reached.

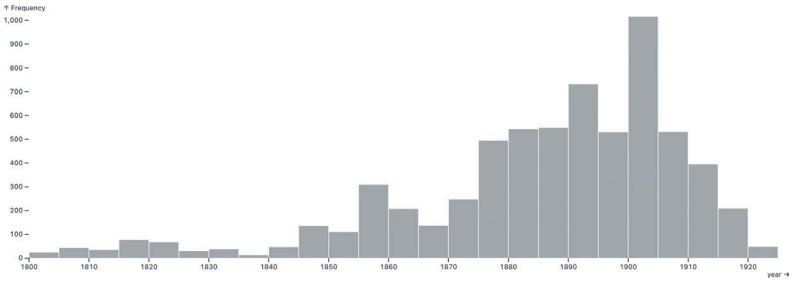


Figure 5.24: Number of articles per year, from 1800 to 1920. Credit: King's Digital Lab.



Figure 5.25: Political leaning of the newspaper titles from which we sourced the articles. Credit: King's Digital Lab.

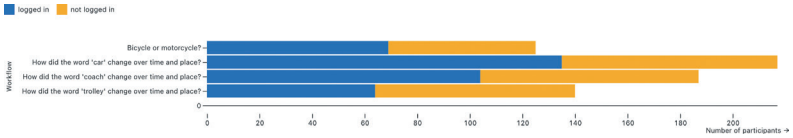


Figure 5.26: Number of annotators that contributed to each workflow. Credit: King's Digital Lab.

While the number of articles for which meanings were annotated vary (371 for 'bike', 1,243 for 'coach', 436 for 'car'), proportionally the higher agreement is found for the word 'coach', where 1,243 articles reach an agreement rate of annotation greater or equal to 65 per cent (i.e. at least 2 out of 3 annotators agree on the meaning associated to the occurrence of 'coach').

Appendix 2 – Design and implementation of visualisation and analysis Notebooks for crowdsourced data

Two of the Notebooks designed and developed by KDL in 2023 are relevant for the analysis of the language of mechanisation discussed above, namely:

1. A GitHub repository with code to produce Jupyter Notebooks to analyse and visualise various crowdsourcing tasks on the

Zooniverse platform to provide insight into the overall annotation activity across workflows.[47] One of the main goals of this work was to make the code used for this project reproducible so that others can use or adapt it to visualise data from their own crowdsourcing projects.

2. An Observable interactive Notebook to analyse words that potentially changed meaning during the nineteenth century, examining their connection to the process of mechanisation in British society, specifically focused on terms used to describe vehicles.[48]

The design and development phase of these Notebooks was concentrated within approximately ten development iterations, grouped into increments and scheduled into a set of two-week periods of development called timeboxes or sprints. In the following section we will describe our

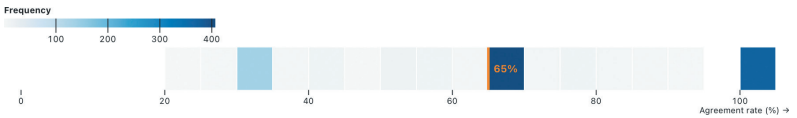


Figure 5.27: Level of agreement around the meaning of the word 'trolley' across the volunteer annotators. Credit: King's Digital Lab.

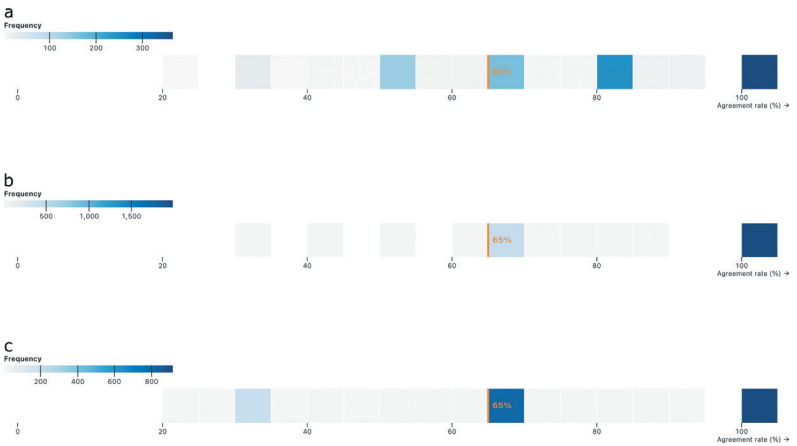


Figure 5.28: Level of inter-annotator agreement around the meaning of the words a) 'car'; b) 'bike' and c) 'coach' respectively. Credit: Barbara McGillivray.

design process that produced these two products and associated visualisations and walk the readers through a series of visualisations to support the analysis.

Technologies and processes

The high-level Zooniverse data model that underpins the crowdsourced datasets fed to the Language of Mechanisation Notebook is structured around subjects (articles/images) and classifications (annotations on the articles/images as described above).

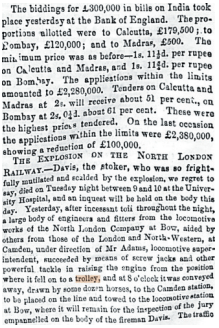
The components of the data model can be summarised as follows:

- workflow: a series of tasks that accomplish a specific purpose on selected subject sets;
- user: a Zooniverse volunteer contributor;
- subject: a media file shown to volunteers during a task. In this case an image of a newspaper article;
- subject_set: groups or 'folders' of subjects;
- annotation: a value assigned to a subject by a volunteer when completing a task in a workflow, for example, a transcription, classification or count;
- comment: all posts on a Zooniverse project Talk board, which can be about specific subjects or general discussion threads;
- tag: hashtags used in comments.

Once data are collected from Zooniverse the following steps are applied to data workflows using tools developed by the Living with Machines team:

- Any duplicates in the subjects are merged.
- The relevant data are selected.
- High-level metadata are aggregated.
- Subjects, annotations and data sources are selected.

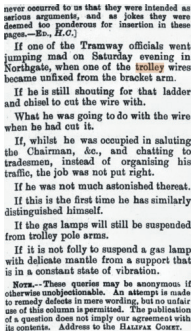
Jupyter Notebooks were chosen by the project team as a technical platform as they are a widely adopted interactive environment that any Zooniverse project manager could use as a starting point to visualise information about their crowdsourcing workflows. Our initial Notebooks were later extended to prepare and export data for further data visualisation and analysis in other platforms.



The biddings for £300,000 in bills on India took place yesterday at the Bank of England. The proportions allotted were to Calcutta, £179,500; to Bombay, £120,000; and to Madras, £500. The minimum price was as before—1s. 11¼d. per rupee on Calcutta and Madras, and 1s. 11½d. per rupee on Bombay. The applications within the limits amounted to £2,280,000. Tenders on Calcutta and Madras at 2s. will receive about 51 per cent., on Bombay at 2s. 0¼d. about 61 per cent. These were the highest prices tendered. On the last occasion the applications within the limits were £2,380,000, showing a reduction of £100,000.

THE EXPLOSION ON THE NORTH LONDON RAILWAY.—Davis, the stoker, who was so frightfully mutilated and scalded by the explosion, we regret to say, died on Tuesday night between 9 and 10 at the University Hospital, and an inquest will be held on the body this day. Yesterday, after incessant toil throughout the night, a large body of engineers and fitters from the locomotive works of the North London Company at Bow, aided by others from those of the London and North-Western, at Camden, under direction of Mr Adams, locomotive superintendent, succeeded by means of screw jacks and other powerful tackle in raising the engine from the position where it fell on to a trolley, and at 8 o'clock it was conveyed away, drawn by some dozen horses, to the Camden station, to be placed on the line and towed to the locomotive station at Bow, where it will remain for the inspection of the jury empannelled on the body of the fireman Davis. The traffic

Figure 5.29: Example of a 'subject' – an image taken from an 1864 article where the word 'trolley' is highlighted. In this case Zooniverse volunteers annotated it with the classification of 'a cart, or wheeled stand or table used to serve, transport or sell food, objects, luggage, etc., powered by people or animals'. Credit: Mia Ridge; British Library.



never occurred to us that they were intended as serious arguments, and as jokes they were deemed too ponderous for insertion in these pages.—Ed., *H.C.*)

If one of the Tramway officials went jumping mad on Saturday evening in Northgate, when one of the trolley wires became unfixed from the bracket arm.

If he is still shouting for that ladder and chisel to cut the wire with.

What he was going to do with the wire when he had cut it.

If, whilst he was occupied in saluting the Chairman, &c., and chatting to tradesmen, instead of organising his traffic, the job was not put right.

If he was not much astonished thereat.

If this is the first time he has similarly distinguished himself.

If the gas lamps will still be suspended from trolley pole arms.

If it is not folly to suspend a gas lamp with delicate mantle from a support that is in a constant state of vibration.

NOTE.—These queries may be anonymous if otherwise unobjectionable. An attempt is made to remedy defects in mere wording, but no unfair use of this column is permitted. The publication of a question does not imply our agreement with its contents. Address to the HALIFAX COMET.

Figure 5.30: Another subject, an image of part of an article from 1903 where the word 'trolley' is highlighted. In this case Zooniverse volunteers annotated it with the classification of 'a trolley-car powered by steam or electricity'. Credit: Mia Ridge; British Library.

Observable was selected as the platform to deliver the Language of Mechanisation visualisations because it offered a powerful platform for data visualisation and exploration that enabled collaborative and rapid prototyping. For example, team members could work together iteratively developing visualisations and sharing insights in real-time. Its interactive nature allowed for experimentation and quick iteration, facilitating the exploration of Living with Machines' complex datasets.

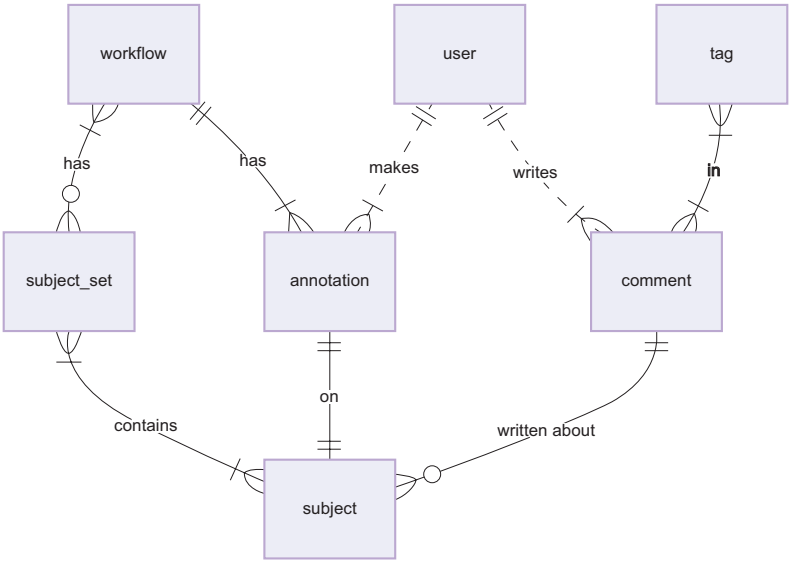


Figure 5.31: Simplified data model of Zooniverse data. A more granular data model includes additional data points which determine the parameters used to produce the visualisations and that the user of the Notebooks can control interactively. Credit: King's Digital Lab.

Additionally, Observable's integration of code and visuals, along with its Notebook-style interface, aligned with KDL and project values in terms of transparency, reproducibility and documentation, as well as project objectives, making the material suitable for data analysis, storytelling and knowledge sharing.

The design process to deliver the products for this project derives from the double diamond process.[49] It consists of four phases with relatively quick iterations within each phase.

Phase 1 – discover

Part of the discovery phase included some user research – supported by open-ended questionnaires – to reach an agreed understanding of the project scope, objectives and audiences as well as for KDL team to gain familiarisation with the data points of the crowdsourced datasets inclusive of newspapers metadata. The main results of this phase were: (1) the definition of a high-level data model that would underpin interactions

with the datasets by potential users; (2) analysis work that informed the following phases.

Phase 2 – define

The outcomes of the discovery phase were integrated with the project requirements (defined at high-level prior to the project starting) and tasks to guide development were created or integrated as needed in an iterative fashion following review of each part of the technical solution (also called an increment). A Miro board was used to support this phase and guide future design iterations.[50]

Phase 3 – develop

A range of different technical solutions – from off-the-shelf plotting libraries to custom-built self-hosted sites – were initially discussed and tested; however, given the time and budget at our disposal, the options quickly converged to adopt existing Notebook platforms (Jupyter and Observable respectively). With the platforms default design system in place, there were obvious user interface constraints and customisation of these platforms would have not been practical; the focus of our work shifted to improving the user experience. Considerable effort went into organising the narrative around the visualisations, improving visual elements, accessibility and adding explanatory text. We explored and tested different chart types and interactions and aimed for a simplified terminology and language for what was possible, to encourage engagement and facilitate understanding.

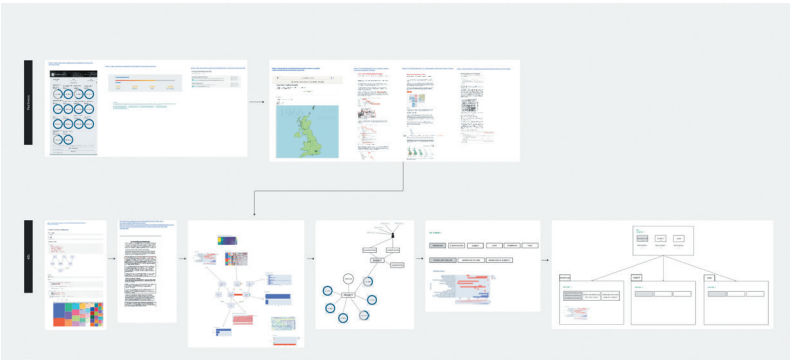


Figure 5.32: Screenshot of static mock-ups sketched in Miro board during the Discovery phase. Credit: King's Digital Lab.

Testing and drafting a statement for digital accessibility was also part of phase 3. The design and development of the interactive Notebook was a very small part of the overall project; the choice of the Observable platform for this imposed limitations which meant that a full accessibility assessment could not be performed on the page. However, KDL and the Living with Machines project both prioritised digital accessibility in their work to enable as many people as possible to access the information, and we set aside resources to improve the issues we found. Some limitations we could not overcome are as follows:

- Visual limitation: most sections of the Notebook were labelled but some charts include dimensions or data elements that are defined only by colour, causing potential difficulty for users with visual colour impairment.
- Device limitation: the Notebook is responsive (i.e. adapts well to mobile as well as desktop devices) but it is better viewed on larger screens as some of the charts might not display or could be hard to view due to the lack of space.
- Browser issues: feedback suggested that some of the charts do not load on certain browsers; the table of contents links do not work in Safari browser.

These issues were noted in the accessibility statement of the Observable Notebook product with a call for user feedback to raise any other issues that would be identified.

Phase 4 – deploy

Adjustments were iteratively made in this phase based on a pragmatic selection of the feedback given by the users interacting with the Notebook.

Near the end of the Develop phase, KDL and the research team set up a quick usability test to stay within budget constraints. KDL prepared a set of questions for test participants to answer – partners gathered participants and conducted the usability test. There were two separate tests for the two outputs. Both had about eight to ten participants recruited from the British Library and wider Living with Machines team. There were no set tasks and the questions were open-ended with the aim to get quick feedback to identify top issues. KDL collated the responses and gave suggestions

to the research team on what we found were the top issues and worked together to implement those improvements.

Notes

1 This chapter presents the work done in the 'Language of mechanisation' work package, one of the main components of the Living with Machines project and led by McGillivray. McGillivray conceived the scope of the work reported on in this chapter and led the design and implementation of the research and the writing process; she also drafted and edited the first and last sections of this chapter. As typical for any project KDL undertakes, the design of the visualisations discussed under 'Visualisations and Analysis' was conducted collaboratively with members of the Living with Machines project (McGillivray, Pedrazzini, Ridge, Kalle Westerling) and coordinated by KDL sub-team of Research Software (RS) Analyst (Ciula), RS Engineer (Vieira), and UI/UX Designer (Ong). Other KDL team members contributed to management and administration (Lab Manager Pam Mellen). Pedrazzini drafted the 'Automatic Detection of Meaning Change in the Lexicon of Mechanisation' section and co-led (with McGillivray) the experiments described therein; Ridge conceived the scope and led the design and implementation of the novel crowdsourced annotation tasks and wrote 'Designing Sense Disambiguation Tasks for Crowdsourcing Volunteers'; Lawrence acted as historical adviser and co-wrote historical interpretations in 'Visualisations and Analysis'. This chapter uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

2 M. Görlach, *English in Nineteenth-Century England* (Cambridge University Press, 1999); C. Kay and K. L. Allan, *English Historical Semantics* (Edinburgh University Press, 2015).

3 Francis Spufford and Jenny Uglow, *Cultural Babbage: Technology, Time and Invention* (Faber and Faber, 1996); Herbert Sussman, *Victorian Technology: Invention, Innovation, and the Rise of the Machine* (Praeger, 2009).

4 Görlach, *English in Nineteenth-Century England*, 10–25; David F. Mitch, *The Rise of Popular Literacy in Victorian England: The Influence of Private Choice and Public Policy* (University of Pennsylvania Press, 1991); David Vincent, *Literacy and Popular Culture: England, 1750–1914* (Cambridge University Press, 1989).

5 Martin Hewitt, *The Dawn of the Cheap Press in Victorian Britain: The End of the 'Taxes on Knowledge', 1849–1869* (Bloomsbury, 2014); Emma Griffin, 'The Making of the Chartists: Popular Politics and Working-Class Autobiography in Early Victorian Britain', *English Historical Review* 129, no. 538 (2014): 578–605, https://doi-org.libproxy.york.ac.uk/10.1093/ehr/ceu078; Simon Eliot, 'Circulating Libraries in the Victorian Age and After', in *The Cambridge History of Libraries in Britain and Ireland*, vol. 3, *1850–2000*, ed. Alistair Black and Peter Hoare (Cambridge University Press, 2006), 125–46.

6 Görlach, *English in Nineteenth-Century England*, 92ff.; *English Historical Linguistics*, ed. Alexander Bergs and Laurel J. Brinton (De Gruyter Mouton, 2012), 1063ff.; Kay and Allan, *English Historical Semantics*, 28ff.

7 Görlach, *English in Nineteenth-Century England*, 125ff.

8 Herbert Sussman, *Victorians and the Machine: The Literary Response to Technology* (Harvard University Press, 1969); Colin Manlove, 'Charles Kingsley, H. G. Wells, and the Machine in Victorian Fiction', *Nineteenth-Century Literature* 48, no. 2 (1993): 212–39; Tamara Ketabgian, *The Lives of Machines: The Industrial Imaginary in Victorian Literature and Culture* (University of Michigan Press, 2011).

9 Bergs and Brinton, *English Historical Linguistics*, 1063ff.

10 Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė Oleškevičienė and Marieke van Erp, 'LL(O)D and NLP Perspectives on Semantic Change for Humanities Research', *Semantic Web* 13, no. 6 (2022).

11 Cf. e.g. Charlotte Brewer, 'The Electronification of the *Oxford English Dictionary*', *Dictionaries: Journal of the Dictionary Society of North America* 25 (2004): 1–43.

12 See Barbara McGillivray, *How to Use Word Embeddings for Natural Language Processing* (SAGE Research Methods: Doing Research Online, 2022), https://doi.org/10.4135/9781529609578, for a non-technical introduction to this technique. See Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu and Simon Hengchen, eds, *Computational Approaches to Semantic Change* (Language Science Press, 2021) for an overview of recent NLP research on semantic change detection.

13 Discussed further in Mia Ridge, Nilo Pedrazzini, Miguel Vieira, Arianna Ciula and Barbara McGillivray, 'Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project', *Journal of Open Humanities Data* 10, no. 1 (2024): 1–9, https://doi-org.libproxy.york.ac.uk/10.5334/johd.195.

14 Raphael Samuel, 'Workshop of the World: Steam Power and Hand Technology in Mid-Victorian Britain', *History Workshop Journal* 3 (1977): 6–72; Maxine Berg and Pat Hudson, 'Rehabilitating the Industrial Revolution', *Economic History Review* 45, no.1 (1992): 24–50.

15 Görlach, *English in Nineteenth-Century England*.

16 See Giorgio Tolfo, Olivia Vane, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, David Beavan and Katherine McDonough, 'Hunting for Treasure: Living with Machines and the British Library Newspaper Collection', in *Digitised Newspapers: A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*, ed. Estelle Bunout, Maud Ehrmann and Frédéric Clavert (De Gruyter Oldenbourg, 2022), 23–46. For a breakdown of the newspaper titles included in this corpus, their temporal coverage and the number of tokens, see Mia Ridge and Nilo Pedrazzini, *Public Domain Newspaper Titles in Living with Machines* (Zenodo, 2024), accessed 4 February 2026, https://doi.org/10.5281/zenodo.18480866.

17 Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, 'Efficient Estimation of Word Representations in Vector Space', *arXiv* (2013), https://doi.org/10.48550/arXiv.1301.3781. We used the implementation of Word2Vec by Gensim: Radim Řehůřek and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (European Language Resources Association, 2010), 45–50.

18 William L. Hamilton, Jure Leskovec and Dan Jurafsky, 'Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2016), 1489–501.

19 Because the models are trained separately, they are not initially comparable with one another, that is, they only make sense 'internally'. For example, we can compare how similar the vectors for 'king' and 'queen' are in the 1860s, but we cannot compare 'king' in the 1860s and 'queen' in the 1890s, or 'king' in the 1840s and 'king' in the 1870s. The vector assigned to each word at the end of a model training is exclusively computed based on the text from the relevant subcorpus (e.g. 1860s, 1870s, etc.), namely on the words occurring there, their frequency and their co-occurrence. The alignment of models is typically achieved by applying an algorithm (Orthogonal Procrustes) which transforms the vectors of a model so that they are comparable to those of other models, while also ensuring that the relationships between words within the original model are preserved. For a more technical explanation on the alignment of diachronic word embedding models, see Nilo Pedrazzini and Barbara McGillivray, 'Machines in the Media: Semantic Change in the Lexicon of Mechanization in 19th-century British Newspapers', in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities* (Association for Computational Linguistics, 2022), 85–95.

20 The decade-level embedding models are openly available in Zenodo: Nilo Pedrazzini and Barbara McGillivray, 'Diachronic Word Embeddings from 19th-century Newspapers Digitised by the British Library (1800–1919)' (Zenodo, 2022), accessed 4 February 2026, https://doi.org/10.5281/zenodo.7181682.

21 We did this with the help of off-the-shelf Python packages based on Levenshtein distance (pyspellchecker).

22 Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray and Giovanni Colavizza, 'Assessing the Impact of OCR Quality on Downstream NLP Tasks', in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, vol. 1, *ARTIDIGH* (2020): 484–96, https://doi.org/10.5220/0009169004840496.

23 Pedrazzini and McGillivray, 'Machines in the Media'.

24 Roberta Killick, Paul Fearnhead and Idris A. Eckley, 'Optimal Detection of Changepoints with a Linear Computational Cost', *Journal of the American Statistical Association* 107, no. 500 (2012): 1590–98.

25 H. Sakoe and S. Chiba, 'Dynamic Programming Algorithm Optimization for Spoken Word Recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, no. 1 (1978): 43–9, https://doi.org/10.1109/TASSP.1978.1163055.

26 Kaspar Beelen, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitised Newspaper Collections: Introducing the Environmental Scan', *Digital Scholarship in the Humanities* 38, no. 1 (2023): 1–22, https://doi.org/10.1093/llc/fqac037.

27 For a fuller account of the linked metadata, see Kalle Westerling, Kaspar Beelen, Tim Hobson, Katherine McDonough, Nilo Pedrazzini, Daniel C.S. Wilson and Ruth Ahnert, 'LwMDB: Open Metadata for Digitised Historical Newspapers from British Library Collections', *Journal of Open Humanities Data* 11, no. 1 (2025): 32.

28 Sarah J. Leslie, Sangeet Khemlan, Sandeep Prasada and Sam Glucksberg, 'Conceptual and Linguistic Distinctions between Singular and Plural Generics', in *Proceedings of the 31st Annual Cognitive Science Society* (2009): 479–84; Alda Mari, Claire Beyssade and Fabio Del Prete, *Genericity* (Oxford University Press, 2012).

29 Pedrazzini and McGillivray, 'Machines in the Media'.

30 Zooniverse, https://www.zooniverse.org, is the world's largest platform for digital volunteering on citizen science and citizen humanities projects, with nearly 3 million registered volunteers. Earlier work with dictionary senses and 'machine' words included two versions of a 'What's that machine? Classify it!' workflow. We are grateful to the hundreds of volunteers whose participation in and comments on these tasks informed the work described in this chapter.

31 Mia Ridge, 'From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing', *Curator: The Museum Journal* 56, no. 4 (2013): 435–50, https://doi.org/10.1111/cura.12046.

32 Mia Ridge, 'What Is a "Machine" Anyway? Help Us Describe Them', *Living with Machines* (blog), 27 October 2020, accessed 24 February 2026, https://livingwithmachines.ac.uk/what-is-a-machine-anyway-help-us-find-out/. Mia Ridge, 'From Prams to Parliament: What Was a Machine? Help Us Find Out', *Living with Machines* (blog), 4 December 2020, accessed 24 February 2026, https://livingwithmachines.ac.uk/from-prams-to-parliament-what-was-a-machine-help-us-find-out/.

33 Mia Ridge, Nilo Pedrazzini and Zooniverse volunteers, 'What "Machine" Words Changed When?', *Living with Machines Talk*, January 2023, https://www.zooniverse.org/projects/bldigital/living-with-machines/talk/2795/2773552?comment=4555192&page=1 (accessed 25 June 2026).

34 Mia Ridge, 'Outreach and Marketing for Crowdsourcing Tasks', *Living with Machines* (blog), 27 June 2024, accessed 24 February 2026, https://livingwithmachines.ac.uk/outreach-and-marketing-for-crowdsourcing-tasks/.

35 Mia Ridge, Nilo Pedrazzini, Miguel Vieira, Arianna Ciula and Barbara McGillivray, 'Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project', *Journal of Open Humanities Data* 10, no. 1 (29 April 2024): 33, https://doi.org/10.5334/johd.195. The datasets are: Language of Mechanisation – annotated historical newspaper articles, https://doi.org

/10.23636/5t9m-0g59; and OCR and crowdsourced annotations, Language of Mechanisation, JSON files https://doi.org/10.23636/z634-km37.

36 *Oxford English Dictionary*, s.v. 'trolley (*n.*)', March 2024, https://doi.org/10.1093/OED/2298568645.

37 Oxford English Dictionary, s.v. 'bicycle (n.)', December 2023, https://doi.org/10.1093/OED/4063130776.

38 Oxford English Dictionary, s.v. 'bike (n.2)', March 2024, https://doi.org/10.1093/OED/1131314560, Glen Norcliffe, *Ride to Modernity: The Bicycle in Canada, 1869–1900* (University of Toronto Press, 2001), 47–58.

39 Ryan Cordell, David A. Smith, Abby Mullen, Jonathan D. Fitzgerald and Avery Blankenship, *Going the Rounds: Virality in Nineteenth-Century American Newspapers* (University of Minnesota Press, Manifold Platform, 2025), https://manifold.umn.edu/projects/going-the-rounds (accessed 25 June 2026), which draws on the Viral Texts Project at Northeastern University; Max Franke, Markus John, Moritz Knabben, Jana Keck, Tanja Blascheck and Steffen Koch, 'LilyPads: Exploring the Spatiotemporal Dissemination of Historical Newspaper Articles', *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (VISIGRAPP 2020), 3 (IVAPP): 17–28, which comes out of the Oceanic Exchanges project: https://oceanicexchanges.org/ (accessed 25 June 2026), https://doi.org/10.17605/OSF.IO/WA94S

40 Oxford English Dictionary, s.v. 'coach (n. & adv.)', March 2024, https://doi.org/10.1093/OED/2505898911.

41 Oxford English Dictionary, s.v. 'car (n.1)', February 2024, https://doi.org/10.1093/OED/4164178241.

42 Oxford English Dictionary, s.v. 'coach (n. & adv.)', March 2024, https://doi.org/10.1093/OED/2505898911.

43 Oxford English Dictionary, s.v. 'car (n.1)', February 2024, https://doi.org/10.1093/OED/4164178241.

44 David V. Herlihy, *Bicycle: The History* (Yale University Press, 2004), 159–282.

45 David Kimberley, *Coventry's Motorcycle Heritage* (History Press, 2009).

46 Chris A. Williams, 'Risk on the Roads: Police, Motor Traffic and the Management of Space, c. 1900–1950', in *Governing Risks in Modern Britain: Danger, Safety and Accidents, c. 1800–2000*, ed. Tom Crook and Mike Esbester (Palgrave Macmillan, 2016), 195–219.

47 King's Digital Lab, 'Data Analysis and Visualisation of Zooniverse Crowdsourcing Tasks', kingsdigitallab/lwm-davizct, 13 April 2023, accessed 6 February 2026, https://github.com/kingsdigitallab/lwm-davizct.

48 Miguel Vieira, Tiffany Ong, Arianna Ciula, Barbara McGillivray, Mia Ridge and Nilo Pedrazzini, 'Language of Mechanisation', Observable, 4 March 2025, accessed 6 February 2026, https://observablehq.com/@jmiguelv/language-of-mechanisation.

49 See ‘The Double Diamond’, Design Council, accessed 6 February 2026, https://www.designcouncil.org.uk/our-work/skills-learning/the-double-diamond/.

50 Miro, accessed 6 February 2026, https://miro.com.

Afterword: historians and the future of AI

Ruth Ahnert

As I write this afterword in January 2026, Living with Machines feels, in many ways, like a project ahead of its time. ChatGPT was launched during the last few months of the project; since then, amid both the hype and the catastrophising about AI, a growing chorus of voices has called for deeper engagement with the humanities and social sciences. The aim of this Afterword is briefly to outline the substance and character of these proposals, and to examine the extent to which such future-oriented provocations found early, partial answers in our own experimental interdisciplinary project – nestled within a data science and AI institute, housed in the UK's national library. The lessons that the Living with Machines project offers to existing interlocutors in this space, as well as to new collaborative teams now entering the fray, arise not only from the outcomes of our research but also from the methods, assumptions and commitments that shaped our approach to interdisciplinary collaboration itself.

Calls for engagement from the humanities and social sciences in the development of AI grew significantly in 2025. Several influential publications and funding initiatives that emerged during that year have generated both a sense of collective momentum and renewed urgency. Earlier in 2025 Henry Farrell, Alison Gopnik, James Evans and Cosma Shalizi published a position paper arguing that 'large AI models should not be viewed primarily as intelligent agents but as a new kind of cultural and social technology'.[1] At its core, the paper constitutes a call for interdisciplinary collaboration:

> It would be far easier for social scientists and computer scientists to cooperate and combine their respective strengths if both understood that [language models] are no more – but also no less – than a new kind of cultural and social technology. Computer scientists could bring together their deep understanding of how these systems work with social scientists' comprehension of how other such large-scale systems have reshaped society, politics, and the economy in previous eras, elaborating existing research agendas and discovering new ones. This would help remedy past confusions in which computer scientists have adopted overly simplified notions of complex social phenomena while social scientists have failed to understand the complex functioning of these new technologies.

The weight these authors bring in terms of their own interdisciplinarity and experience of collaboration – as scholars from the fields of political science, developmental psychology and cognitive science, computational social science and predictive modelling – makes a compelling case for the benefits of such cooperation.

Written in a similar mode of provocation, but instead making the case for the need for a specifically *humanities* perspectives, is a 2025 pre-print authored by Lauren F. Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton and David Mimno – a group that contains leading figures from the computational humanities with roots both in humanities and computer science departments. The significance of this paper lies in its articulation of a series of concrete, manifesto-like values and contentions, reminiscent of Klein's influential co-authored monograph *Data Feminism*:

1) Models make words, but people make meaning;
2) Generative AI requires an expanded definition of culture;
3) Generative AI can never be representative;
4) Bigger models are not always better models;
5) Not all training data is equivalent;
6) Openness is not an easy fix;
7) Limited access to compute enables corporate capture; and
8) AI universalism creates narrow human subjects.[2]

Under these organising statements, the authors outline not only the critiques that a humanities-oriented perspective brings to the design

and practice of generative AI but also a set of constructive proposals for how such complexities might be navigated by combining data-driven approaches with humanistic epistemologies and expertise.

Crucially, these calls for AI-humanities collaboration are not just coming from small collectives of interdisciplinary academics. Funders, too, through targeted calls and initiatives are actively seeking to open up a space – perhaps even a new disciplinary formation – at the interface of the humanities and AI. Significantly, calls are coming from funders both in the humanities and the sciences. In 2022 the UK's Arts and Humanities Research Council (AHRC) funded the Bridging Responsible AI Divides (BRAID) programme, which seeks to embed arts and humanities research into the responsible AI ecosystem. The programme continues to drive a cultural shift through the delivery of demonstrator projects and fellowships. Building on this momentum, in 2025 the AHRC, in partnership with Canada's Social Sciences and Humanities Research Council, invited expressions of interest to attend 'a humanities-led, interdisciplinary research sandpit putting humanities insights and methodologies at the heart of artificial intelligence (AI) tech design'.[3] The sandpit was to be informed by the 'Doing AI Differently' White Paper, whose cross-disciplinary list of co-authors call for 'a fundamental shift in AI development – one that positions the humanities, arts and qualitative social sciences as integral, rather than supplemental, to technical innovation'.[4]

A similar interdisciplinary orientation underpinned a further funding call, also announced in 2025, by Schmidt Sciences, a US-based philanthropy established by Eric and Wendy Schmidt to support unconventional research across science and technology. Through its new Humanities and AI Virtual Institute (HAVI), Schmidt Sciences issued a call focused on 'bridging the gap between humanities scholarship and AI capabilities'. The call explicitly framed AI as offering new opportunities for humanistic research, while emphasising how humanistic epistemologies and expertise 'can significantly advance and improve AI technology'.[5] As noted in the press release announcing the first round of HAVI awards in December 2025, most general-purpose AI models are ill-suited to humanities research because they are trained primarily on contemporary data, modern languages and two-dimensional media. By contrast, humanities research often engages with 'ancient or lesser-spoken languages, three-dimensional artefacts, art made from a variety of materials, and relatively small amounts of ambiguous and culture-specific information'.[6] The funded projects – including one

that grew out of Living with Machines, discussed further below – seek to address these challenges by developing AI systems capable of answering questions from specific historical or geographical perspectives; supporting the digital reconstruction and interpretation of lost, damaged or dispersed artefacts; and applying multimodal and technical integration to the study of art, music and scientific imagery. Many foreground interpretability, provenance and human-in-the-loop collaboration, emphasising how AI can support rather than supplant human expertise, as well as proposing frameworks for critically assessing AI outputs and integrating ethical considerations into system design.

The Living with Machines project anticipated some of these current developments both in terms of the methodological approaches we brought to historical questions and of the collaborative modes through which we sought to address them. Of course, as the previous chapters have made clear, in many cases our work on the project did not involve AI because it did not require that level of complex and resource-demanding pipelines. We would be the first to say that data science *without AI* might be the most pertinent approach for the majority of historical questions. However, in what follows, I suggest five ways in which both the work and practice on Living with Machines can inform the emerging wave of scholarship at the interface of the humanities and AI. These serve both as a synthesis of what we have sought to demonstrate through this book, and as a set of provocations grounded in hard-won experience, that can be added to those now circulating in this research space.

No analysis without provenance

The most important thing we can do for the future of AI is to attend to the nature of our data. As Klein and colleagues observe: 'recognition of the cultural complexity of historical sources, including training data, [. . .] offers a fundamentally different way to approach the issue of *bias*' (emphasis mine).[7] This recognition underpinned the development of the Environmental Scan on our project, both a method and framework for source criticism at scale. Our contention was that, if historians' established practices of source criticism are to be scaled up, they require a computational approach. By identifying contextual data sources – Mitchell's press directories, in our case – we demonstrated how the Environmental Scan

offers two complementary approaches to examining bias in digitised newspaper datasets.[8] First, the *wide scan* assesses how much of the potential nineteenth-century newspaper landscape is represented within a given digitised corpus. Second, the *deep scan* uses the contextual source to generate new metadata fields for each title, describing key attributes of our titles – such as frequency of publication, area of circulation, price and self-declared political leaning – allowing us to evaluate representativeness across multiple dimensions of the data. In doing so, we were already undertaking the work that Klein and colleagues describe in their call for 'rich metadata [that] can help to ensure that datasets more accurately represent the contents they seek to include'.[9]

At the same time, by foregrounding the *multidimensionality* of representativeness, the work on Living with Machines reinforces the observation (also made by Klein et al.) that neither data nor the AI systems trained upon it can ever be fully representative. Our aim should not be to 'de-bias' data because this is an impossible goal.[10] Rather, if we look at the problem through feminist and intersectional approaches to archives and data, we instead emphasise the importance of making biases and archival silences visible.[11] Our motivation should be to identify the dimensions of our data that shape the results of data-driven analysis, in order to inform future data practices, and to develop methods that explicitly account for those influences.[12] Historians, of course, remain constrained by what survives in the archive, but nonetheless have agency. We can supplement datasets with additional sources; we can make principled decisions about which archives are prioritised for digitisation; and we can design analytical samples that reflect our research values. In the case of nineteenth-century newspapers, for example, methods such as the Environmental Scan can guide the prioritisation of titles representing communities marginalised by class or geography. Where further digitisation is not feasible – a growing reality across the humanities and cultural heritage sector – we can construct custom subsamples, oversampling marginalised voices and downsampling those associated with dominant groups. Such strategies are not attempts to 'correct' bias, but should be seen rather as experimental interventions, designed to explore how structural inequalities shape computational interpretations of the past, and to ask what histories emerge when marginalised perspectives are amplified.

Methods and models must be situated

This notion of viewing history from a specific, marginalised perspective leads to my second provocation: because human knowledge is inherently situated, the methods and models developed in AI research must be too. This was a central concern of Living with Machines from its inception, as we grappled with the problem of how computational methods might account for historical specificity. A growing body of scholarship has since articulated a core limitation of generative AI: its difficulty in accounting for situated knowledge, whether that be particular geographies, historical moments, cultural contexts or aesthetic traditions. This concern has clear parallels within AI research itself, where homogenisation is increasingly recognised as a threat not only to diversity of expression but also to accuracy and equity.[13] As Edwin Rowland, Richard Jean So and Hoyt Long put it, 'Monoculture exercises social power when it reproduces already dominant worldviews and amplifies existing inequalities.'[14] Systems that cannot accommodate the perspectives of different communities and user groups risk producing results that are at best misleading and at worst actively harmful when deployed in real-world contexts.[15]

One source of this homogenisation lies in the epistemological assumptions embedded in contemporary AI systems. As briefly discussed in the introduction, the neo-Kantian philosopher Wilhelm Windelband famously distinguished between two forms of knowledge: the *nomothetic* and the *idiographic*. Nomothetic approaches, which he associated with the sciences, aim to identify general laws and abstractions. Idiographic approaches, which he linked to the humanities, focus instead on the particular, the contingent and the historically or subjectively situated.[16] While this distinction inevitably oversimplifies the relationship between the sciences and the humanities, it is nonetheless useful for understanding both the trajectory of AI development and its limitations. Generative models inherit a nomothetic approach to knowledge production, often at the expense of situated knowledge. Large language models (LLMs) excel at identifying general patterns across vast corpora, but they struggle to preserve what historians care about most: change over time, contextual nuance and the meanings that emerge from specific historical circumstances.

That this remains a challenge is perhaps surprising, given that the field of historical linguistics has long used computational approaches, and that semantic change detection has become an active subfield in the last

decade.[17] Chapter 5 of this book offers an example of work in this space, tracing how the meanings of words such as *gear*, *car*, *train*, *railway* and *trolley* shifted in response to the rapidly changing technological landscape of nineteenth-century Britain. These are not static terms but historically contingent ones, whose meanings cannot be recovered without attention to temporal context. Computational linguists have also been at the helm of projects developing historically situated language models, such as work on temporally segmented datasets, and the training of language models exclusively on nineteenth-century texts for developing approaches to historical reasoning.[18]

There is a growing consensus that the development of contextualised models is urgent. Yet Living with Machines was already engaged in this work before the eruption of work on LLMs following ChatGPT's release in late 2022. From an early stage the project explored how language models could be trained to incorporate historical context directly, for example by conditioning language models on metadata such as year and place of publication.[19] By incorporating this information, we enabled our models to associate linguistic patterns with historical moments. In practical terms, this meant that the model learned not only which words tend to co-occur but *when* and *where* particular meanings were likely to hold. This approach anticipated recent calls for smaller, more carefully curated models trained for specific historical tasks, rather than ever-larger general-purpose systems.[20]

Our emphasis on situatedness also shaped our investigation of what we termed *atypical animacy*: the attribution of human qualities to machines, and the mechanisation of humans, in nineteenth-century texts. After fine-tuning a language model on defined periods within a corpus of around 60,000 nineteenth-century books, we asked the model to predict the missing word in sentences where we had masked machine-related words.[21] Because language models tend to prefer conventional usage, moments where the model predicted animate entities (such as *man*, *woman*, *child, prince* or *cow*) instead of *machine* showed us that these sentences were attributing animacy, intention or emotion to the machine. This method outperformed previous computational approaches to detecting atypical instances of animacy.[22] Perhaps more importantly, it also demonstrated the importance of context in identifying figurative and innovative language use: closer inspection of the model's predictions revealed that they reflected the biases, assumptions and attitudes of the

historical material on which the model had been trained.[23] In other words, the model did not produce an abstract or universal account of animacy but one that was historically situated, thereby showing how models can learn the cultural and temporal contexts of their data.

While promising work is beginning to emerge in this space, much more research is needed; and this is where the spin-off project Text Machine: Computing Literary Innovation (funded as part of the first HAVI round) picks up the baton.[24] The project itself is a study of how innovation in linguistic usage relates to moments of technological change, and proposes to capitalise on language models' tendency towards normative patterns of language in order to find moments where the model is 'surprised'. To understand the temporal specificity of such moments of surprise, we will first fine-tune our model on different time slices of English literature over the period 1473 to the present. Our prior experience on the Living with Machines project suggests that attending to situatedness is not a constraint on computational research, but a means of making it more historically aware, methodologically rigorous and culturally sensitive. For AI research, this means new capabilities in temporal reasoning, contextual adaptation and nuanced language understanding across domains and languages. But such work requires more than just historical data; it needs historical expertise to interpret the relationship between the outputs of such approaches to modelling, and the *inputs* – that is to say, we need expertise in the sources and the contexts in which they were created. Without such expertise it is too easy to come to false conclusions.

Knowledge is inherently multimodal

Recent publications and funding calls have rightly emphasised the need for the development of multimodal AI approaches, noting that many objects of humanistic inquiry – visual art, photography, film, music, performance and material culture – cannot be adequately studied through textual analysis alone.[25] While this recognition marks an important shift, it risks framing multimodality as a problem confined to non-textual sources.[26] Multimodality is not simply a feature of particular kinds of evidence; it is a defining characteristic of human reasoning and historical understanding.

Historians rarely work with a single source type in isolation. Instead, interpretation emerges through the synthesis of heterogeneous materials: texts, images, objects, geo-spatial arrangements, demographic records and

narrative accounts, each offering partial and situated insight. Knowledge is produced not by privileging one modality over others but by holding them in productive tension. This epistemological reality poses a deeper challenge for AI-driven approaches than the question of how to incorporate images or sound alongside text. It asks how computational methods can support forms of reasoning that are layered, contextual and relational.

This challenge was already central to Living with Machines, and particularly Chapter 3, 'Beyond the Tracks', long before multimodality became a prominent concern in AI research. That chapter set out to understand the impact of rail infrastructure on British communities in the nineteenth century – a question that could not be answered through any single dataset or method. Cities, as the chapter notes, are almost irreducibly complex, yet they stand at the heart of human culture and social life, especially in moments of dramatic material change such as industrialisation. Capturing that complexity required a deliberate *convergence* of both sources and methods. It combined geo-coded street-level census data, digitised historic Ordnance Survey maps capturing rail infrastructure, and a longitudinal dataset of railway passenger stations. This was not merely a technical achievement – though linking such sources at a national scale and fine-grained resolution was itself a milestone – but an epistemological one. Structuring historically rich digital sources so that they could be combined in flexible ways made it possible to move between macro-level patterns and street-level social context, aligning quantitative scale with human-scale interpretation.

While one would not characterise the work on Living with Machines as being multimodal per se, this 'convergence experiment' shows how historical understanding depends on the careful orchestration of multiple forms of evidence. We might posit a wider definition of multimodality, not simply as a matter of adding images or audio to text-based systems but as a property of humanistic questions themselves. Progress in this subfield of humanities-focused AI, therefore, depends less on building ever more comprehensive models, and more on historically and culturally informed strategies for combining sources, methods and interpretive frames.

Collaboration is an intellectual practice

Across the recent provocations, position papers and funding calls outlined above, there is a striking degree of consensus: meaningful progress

in AI that harnesses cultural and social data requires deep collaboration with the humanities, social sciences and allied professional communities. What is less often addressed is what such collaboration actually entails in practice. Interdisciplinarity is frequently invoked as an aspiration but rarely examined as a form of work with its own epistemological, organisational and cultural demands.

One of the most significant outcomes of the Living with Machines project was that it functioned as an experiment in what we called 'radical collaboration'.[27] While the project produced substantial scholarly results, I know many of the participants would point to the collaborative experience itself as its most important contribution.[28] As we stated in the introduction of this book, from the outset Living with Machines brought together a diverse team: data scientists, software engineers, historians, digital humanists, curators, library professionals, computational linguists, literary critics and an urban geographer. This was not simply a matter of assembling complementary expertise but of learning how to work across deeply different traditions and institutional realities. These traditions manifest not only in epistemological differences (i.e. between the nomothetic and idiographic approaches, discussed earlier) but also practical and procedural. We needed to counter questions about what counts as a result in different disciplines and professional contexts and how those results should be communicated. Moreover, it is important to note that humanities researchers are rarely trained for collaborative, lab-based research. In this context, questions of authorship, contribution and responsibility proved especially consequential, challenging long-standing norms within the humanities while also exposing unspoken assumptions within computational research.

To navigate these tensions, we found it helpful to think in terms of Peter Galison's trading zones: spaces where different communities can coordinate action without requiring full epistemological alignment.[29] Constructing and maintaining such spaces requires active cultivation, both through cultural and organisational practices, which we discuss at greater length in an earlier output from the project, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project*.[30] In this book, my co-authors and I placed particular emphasis on establishing clear terms of engagement, not only through formal project documentation such as the delivery plan but also through a collaboratively authored project charter, which articulated shared values and expectations about

how the team would work together. Equally important was the creation of a culture of exchange and community. Reading groups – ranging from historiography to Victorian fiction – helped build shared intellectual reference points. Informal practices such as 'coffee and code' sessions lowered barriers between roles and disciplines, while shared training opportunities enabled participants to develop a working literacy in one another's methods. Perhaps most important of all, we got research going by adapting the principle of the minimum viable product from software engineering (turning it into the 'minimum research outcome'), which helped us to articulate the research question of interest for each of the contributing disciplines, to begin work in an efficient and minimal way, and to iteratively develop our questions and approaches.[31]

At an organisational level, the scale of the project required robust systems for managing work, communicating decisions and documenting processes for both the team and our future selves. Here, we again looked beyond the humanities to the established collaborative practices of the software engineers and our colleagues in the British Library, many of which were embraced with enthusiasm by historians once their value became clear. Version control platforms such as GitHub, agile-inspired sprint cycles and an emphasis on documentation all helped support iterative development across disciplinary boundaries. Importantly, documentation was understood broadly: not only as a record of code but as a trace of intellectual, methodological and administrative decisions. Tools such as Jupyter notebooks proved particularly powerful, serving simultaneously as sites of analysis, documentation and collaboration. They enabled team members to work together in ways that made assumptions visible, supported reproducibility and fostered shared understanding across technical and non-technical roles. Similarly, practices of co-authorship – while challenging established norms within the humanities – became a way of acknowledging the collective nature of the work and setting an agenda for more inclusive forms of scholarly credit.

Looking back, what made Living with Machines distinctive was not simply that it brought different disciplines together but that it treated collaboration itself as an object of inquiry and experimentation. The practices that I have briefly described were not ancillary to the research; they were part of the research infrastructure. Progress depended on a willingness to move together while recognising distinct skills and perspectives; on curiosity rather than defensiveness; and on a conscious effort to lower

individual ego in service of collective understanding. If current calls for humanities–AI collaboration are to succeed, they will need to take these lessons seriously. Genuine progress requires not just interdisciplinary teams but the infrastructures, practices and dispositions that allow such teams to think and work together.

Build for reuse, not just release

From its inception, Living with Machines set out not only to answer substantive historical questions but also to develop computational models, tools, code and infrastructure that might be reused, adapted and extended by others. This commitment to transferability was not incidental; it reflected a belief that collaboration does not end with the project team but must extend outward to the wider research community. Experience has taught us, however, that openness alone is insufficient. Making data and code available is a necessary condition for reuse, but it is not a guarantee of it. There is a significant gap between releasing assets at the end of a project and creating data and tools that are genuinely useful, usable and actually used by others.

A helpful way to think about this is in terms of levels:

- Level 1: Available – data and code are openly released, typically at the end of a project.
- Level 2: Usable – assets are documented, licensed permissively and structured with reuse in mind.
- Level 3: Used – assets are taken up, adapted and extended by communities beyond the original team.

Of course, many projects and institutions produce digitised content or research data that never even reaches Level 1, either because of intellectual gatekeeping or, particularly in the case of cultural heritage institutions under financial pressure, through the paywalling of resources.[32] For projects seeking to develop community resources, however, the failure to move beyond Level 1 is often not one of intent but resourcing. Funded research tends to prioritise the delivery of assets rather than their long-term care. When a project ends, so, too, does the costed time of those responsible for maintaining data, code and documentation and therefore the afterlife depends on good will or a sense of personal ownership. Unused data and

code are not just a poor return on investment for funders but a dispiriting outcome for researchers who have invested years of labour in creating resources that struggle to find a life beyond their original context.

These challenges motivated a pilot project on which some of the team participated, and for which Living with Machines served as a core case study. The central premise of Data/Culture: Building Sustainable Communities around Arts and Humanities datasets and software (2023–25) was simple: reuse does not happen automatically; it must be supported, curated and sustained. MapReader, the software we discussed in Chapter 2, was a key focus of the Data/Culture project. Developed initially as part of Living with Machines, thanks to the work enabled over the eighteen months of the Data/Culture project, MapReader now exists as a standalone software library, supported by extensive documentation, open repositories and multiple access points. Data and models are shared via established platforms, outputs have been returned to partner institutions such as the National Library of Scotland, and training workshops and community calls are being used to lower barriers to entry and cultivate new users.

This experience of working with MapReader has reinforced several pieces of practical advice for projects seeking to move beyond release towards reuse. Firstly, reuse should be planned from the start, not as an afterthought at project end. Funders should care as much about open data as they do about open access publishing: it should be a condition of the grant. Secondly, time and resources need to be assigned to the development of the documentation as well as the code and data themselves. This might take the form of data cards, notebooks or data papers. Thirdly, researchers should think about how to aid discoverability within the current data landscape of historical and cultural heritage data, which has, as I have argued with co-authors elsewhere, 'become so fragmented, complicated, and opaque that it hinders both discovery and reuse'.[33] In this article, we suggest various strategies to aid discovery, including community coordination within subfields of the humanities to work towards shared norms for depositing data in the same locations – whether that be national services such as UK Data Service, institutional or heritage repositories, or open platforms like Zenodo. Such coordination could be spearheaded by learned societies and journal policies. But 'depositing data without signposting it is like burying it': data papers, cross-linking through DOIs, indexing via reporting systems and citation in related publications all help ensure that datasets are discoverable and their creators' labour

is recognised.[34] Researchers can also create opportunities for encounters with their data and code through workshops, summer schools and community calls that turn potential users into collaborators. Finally, and above all, sustainable research software and data are not maintained by individuals alone. They are sustained by communities. Building those communities – through care, recognition and ongoing exchange – is work in its own right, and work that the humanities are only beginning to value appropriately. If collaboration is to be taken seriously as a practice, then openness must be understood not as an endpoint but as the beginning of a longer, collective commitment.

Taken together, the provocations in this Afterword argue for a conception of AI-inflected humanities research that is audited, situated, multimodal, collaborative and open in ways that extend beyond the life of individual projects. Living with Machines was never simply an application of computational methods to historical questions but an experiment in how such work might be done differently: across disciplines, scales, sources and forms of expertise, and with sustained attention to infrastructure, care and reuse. Many of the concerns now shaping debates about AI were not fully articulated when this project began, yet they became central to our agenda through practice rather than prescription. If the book has a broader contribution, it lies as much in how the work was organised and shared as in the findings it reports. In that spirit, we close by acknowledging the University of London Press for their willingness to support an experimental open review process, and the wider scholarly community whose engagement helped shape this volume. As both a research project and a collective endeavour, Living with Machines demonstrates that genuine progress in this space depends not on technical innovation alone but on the willingness to rethink how we work, with whom and to what ends.

Notes

1 Henry Farrell, Alison Gopnik, Cosma Shalizi and James Evans, 'Large AI Models Are Cultural and Social Technologies', *Science* 387, no. 6739 (2025): 1153–6 (p. 1156).

2 Lauren F. Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton and David Mimno, 'Provocations from the Humanities for Generative AI Research', *arXiv* (2025): 1, arXiv:2502.19190v1, https://doi.org/10.48550/arXiv.2502.19190.

3 https://www.ukri.org/opportunity/artificial-intelligence-humanities-sandpits-canada-uk-and-us/ (accessed 15 June 2026).

4 Drew Hemment, Cody Kommers et al., 'Doing AI Differently: Rethinking the Foundations of AI via the Humanities', White Paper (Alan Turing Institute, 2025), https://www.turing.ac.uk/news/publications/doing-ai-differently. The paper proposes three priority areas for innovation: (1) Interpretive technologies – AI systems that represent multiple valid perspectives rather than producing monolithic outputs, enabling more nuanced, culturally sensitive reasoning across diverse contexts; (2) Alternative architectures for AI – expanding the AI design space beyond current homogeneous approaches through diverse reasoning paradigms grounded in heterogeneous cognitive, cultural and planetary processes; (3) Human–AI ensembles – developing frameworks for sophisticated, collaborative human–AI systems that strengthen our collective intelligence and enhance rather than replace human capabilities in complex decision-making.

5 https://www.schmidtsciences.org/humanities-and-ai-virtual-institute/ (accessed 15 June 2026).

6 https://www.schmidtsciences.org/havi-2025-announcement/ (accessed 15 June 2026).

7 Klein et al., 'Provocations', 7.

8 See Chapter 1; Kaspar Beelen, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan', *Digital Scholarship in the Humanities* 38, no. 1 (2023): 1–22; and Kaspar Beelen, Jon Lawrence, Katherine McDonough and Daniel C.S. Wilson, 'Whose News? Critical Methods for Assessing Bias in Large Historical Datasets', *Computational Humanities Research* 1 (2025): e8, https://doi.org/10.1017/chr.2025.10007.

9 Klein et al., 'Provocations', 10.

10 Katherine Bode, 'Why You Can't Model Away Bias', *Modern Language Quarterly* 81, no. 1 (2020): 95–124.

11 This is a vast field, but we might look for example to Donna Haraway, 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies* 14, no. 3 (1988): 575–99; Michel-Rolph Trouillot, *Silencing the Past: Power and the Production of History* (Beacon Press, 1995); Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (MIT Press, 2020); Ruha Benjamin, *Race after Technology* (Polity Press, 2019); and Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018).

12 See for example the principles outlined in Klein and D'Ignazio's *Data Feminism*, as well as the 'abolitionist tools' proposed by Benjamin in *Race after Technology*, 187ff. For work on our project assessing the downstream effects of data quality, see Daniel Van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray and Giovanni Colavizza, 'Assessing the Impact of OCR Quality on Downstream NLP Tasks', in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence* (ScitePress, 2020), 484–96.

13 See for example Dhruv Agarwal, Mor Naaman and Aditya Vashistha, 'AI Suggestions Homogenize Writing toward Western Styles and Diminish Cultural Nuances', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025): 1–21; Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis and Chloé Clavel, 'The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text', in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024): 3589–604; Jon Kleinberg and Manish Raghavan, 'Algorithmic Monoculture and Social Welfare', *Proceedings of the National Academy of Sciences* 118, no. 22 (2021): e2018340118.

14 Edwin Roland, Richard Jean So and Hoyt Long, 'The Social AI Author: Modeling Creativity and Distinction in Simulated Cultural Fields', *AI & Society* (2025), https://doi.org/10.1007/s00146-025-02790-0.

15 See Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky and Percy S. Liang, 'Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?', *Advances in Neural Information Processing Systems* 35 (2022): 3663–78.

16 Wilhelm Windelband, 'History and Natural Science', *Theory & Psychology* 8, no. 1 (1998): 5–22. For discussion in the context of digital humanities, see Giorgio Caviglia and Nicole Coleman, 'Idiographic Network Visualizations', *Leonardo* 49, no. 5 (2016): 447.

17 See for example, Xuri Tang, 'A State-of-the-Art of Semantic Change Computation', *Natural Language Engineering* 24, no. 5 (2018): 649–76; Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky and Nina Tahmasebi, 'SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection', *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020* (2020): 1–23; Liwen Hou and David A. Smith, 'Detecting Syntactic Change with Pre-Trained Transformer Models', in *Findings of the Association for Computational Linguistics: EMNLP 2023* (2023), pp. 3564–74; Francesco Periti and Stefano Montanelli, 'Lexical Semantic Change through Large Language Models: A Survey', *ACM Computing Surveys* 56, no. 11 (2024): 1–38; Elisabeth Fittschen, Sabrina Li, Tom Lippincott, Leshem Choshen and Craig Messner, 'Pretraining Language Models for Diachronic Linguistic Change Discovery', *arXiv* (2025): arXiv:2504.05523, https://doi.org/10.48550/arXiv.2504.05523.

18 Fittschen et al., 'Pretraining Language Models'. Their models and datasets can be found at: https://huggingface.co/Hplm (accessed 15 June 2026). See also discussion of Ted Underwood and Matthew Wilkens' project 'Artificial Intelligence for Cultural and Historical Reasoning' in the HAVI press announcement, https://www.schmidtsciences.org/havi-2025-announcement/ (accessed 15 June 2026), and in Klein et al., Provocations', 9.

19 See Kaspar Beelen and Daniel van Strien, 'Metadata Might Make Language Models Better', *arXiv* (2022): arXiv:2211.10086, https://doi.org/10.48550/arXiv.2211.10086; and Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo and Barbara McGillivray, 'When Time Makes Sense: A Historically-Aware Approach to Targeted Sense Disambiguation', in *Findings of the Association for Computational Linguistics: ACL-IJCNLP* (2021): 2751–61.

20 See for example Klein et al., 'Provocations', 9.

21 Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza and Mariona Coll Ardanuy, 'Neural Language Models for Nineteenth-century English', *Journal of Open Humanities Data* 7 (2021): 22.

22 Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel C.S. Wilson and Barbara McGillivray, 'Living Machines: A Study of Atypical Animacy', *Proceedings of the 28th International Conference on Computational Linguistics* (2020): 4534–45.

23 Daniel C.S. Wilson, Mariona Coll Ardanuy, Kaspar Beelen, Barbara McGillivray and Ruth Ahnert, 'The Living Machine: A Computational Approach to the Nineteenth-Century Language of Technology', *Technology and Culture* 64, no. 3 (2023): 875–902.

24 See https://github.com/Text-Machine (accessed 15 June 2026).

25 For instance, the recent open calls published by Schmidt Sciences/HAVI (the Humanities and Artificial Intelligence Virtual Institute) and list of funded projects at https://www.schmidtsciences.org/news/ which include several that address multimodal data, including 'Meeting the Vesuvius Challenge'; 'MakingAI: AI-Driven Integration of "Messy" Data in Technical Art History'; 'Musica Subtilior'; 'From Molecules to Masterpieces'; and 'Seeing through Style: Multimodal Advances for Iconographic Analysis'. Beyond this cohort, there is an emerging field of research into the fine-tuning of multimodal AI models to handle the nuances of historical and cultural heritage data, including, for example, Kaspar Beelen and Natasha Kitcher, 'Heritage Weaver: Classifying, Searching, and Linking Museum Data with Multimodal AI', *Anthology of Computers and the Humanities* 3 (2025): 280–96; and Jia Xing, Wei Ren, Du Lei, Lin Zhao, Xue Qin, Hongdang Shao, Wenjie Li, Yirui Han, Zike Yu, Zheng Xu, Rui Yin, Jiantao Yuan, Jun Wang, Wei Chen, Jun Xu, Xiaoping Zhou, Cheng Yang, Wei Zhou and Binbin Zhou, 'Multimodal AI for Yuan Buddhist Sculpture Chronology and Style', *npj Heritage Science* 13, no. 1 (2025): 443.

26 For example, OCR is, in many ways, a multimodal problem.

27 For a more sustained discussion, see our book Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009175548.

28 Our principles informed the THRIVE project and the AHRC's Responsive Mode Pilot (Mission projects), where similar team-based and convening principles were explicitly tested in 2024.

29 See Peter Galison, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997). For a wider discussion of this concept applied to historical research, see Max Kemman, *Trading Zones of Digital History* (De Gruyter, 2021).

30 See Ahnert et al., *Collaborative Historical Research*.

31 See Tolfo et al., 'The Minimum Research Outcome'.

32 For more on this, see Ruth Ahnert, Katherine McDonough and Daniel C.S. Wilson, 'Show Me the Data: New Practices for Historical Sources', *Transactions of the Royal Historical Society* (2026): 1–18. https://doi.org/10.1017/S0080440126100632.

33 For further discussion of these recommendations, see Ahnert et al., 'Show Me the Data'.

34 Ahnert et al., 'Show Me the Data'.

Bibliography

Datasets

Address GB dataset: J. Rhodes, 'AddressGB: Geo-coded British Census Addresses, 1851–1911', Zenodo, 6 November 2024, at https://doi.org/10.5281/zenodo.10473597.

Address GB evaluation dataset: J. Rhodes, 'AddressGB Manual Evaluation Sample'. Zenodo, 6 November 2024, at https://doi.org/10.5281/zenodo.13770048.

Beyond the Tracks data collection: https://doi.org/10.5281/zenodo.15175157.

Book genre classifier dataset: https://huggingface.co/datasets/TheBritishLibrary/blbooksgenre/viewer/annotated_raw.

GB1900 Gazetteer: https://www.pastplace.org/data/#tabgb1900 (geocoded dataset created by Great Britain Historical GIS Project, University of Portsmouth, the GB1900 partners and volunteers).

I-CeM dataset: K. Schurer and E. Higgs, *Integrated Census Microdata (I-CeM), 1851–1911* [data collection] (UK Data Service, 2020) SN: 7481, http://doi.org/10.5255/UKDA-SN-7481-2.

I-CeM dataset: K. Schurer and E. Higgs, *Integrated Census Microdata (I-CeM) Names and Addresses, 1851–1911: Special Licence Access* [data collection]. 2nd edition (UK Data Service, 2022) SN: 7856, http://doi.org/10.5255/UKDA-SN-7856-2.

I-CeM dataset: J.D. Day, Registration sub-district boundaries for England and Wales 1851–1911 (2016), created for the 'Atlas of Victorian Fertility Decline' project (PI: A.M. Reid), ESRC project ES/L015463/1.

I-CeM dataset: Gill Newton and Robert Bennett. *Record-Linkage of Entrepreneurs in the England and Wales Censuses 1851–91 Using BBCE and I-CeM* (2020), https://doi.org/10.17863/CAM.50178.

I-CeM dataset: Bennett, Robert, Harry Smith, C. Van Lieshout and G.H. Newton, 'Sector and Occupation Classification Data Download for Business Sectors, Occupations and Aggregations of Census Data 1851–1911'. Apollo – University of Cambridge Repository (2018), https://doi.org/10.17863/CAM.26433.

Machines Reading Maps, https://purl.stanford.edu/rc349kh8402 (Rumsey collection data and metadata).
MapReader 'railspace' annotations and outputs: https://github.com/Living-with-machines/MapReader/wiki/GeoHumanities-workshop-in-SIGSPATIAL-2022.
Mitchell's Press Directories dataset (1846–1920): https://bl.iro.bl.uk/concern/datasets/adcef12a-bb3d-40d9-871d-5784022a77e8.
Newspaper dataset: 'Language of Mechanisation: Annotated Historical Newspaper Articles', https://doi.org/10.23636/5t9m-0g59.
Newspaper dataset: 'Language of Mechanisation OCR and crowdsourced annotations, JSON files', https://doi.org/10.23636/z634-km37.
Newspaper dataset word embeddings: Nilo Pedrazzini and Barbara McGillivray, 'Diachronic Word Embeddings from 19th-century Newspapers Digitised by the British Library (1800–1919)' (Zenodo, 2022), https://doi.org/10.5281/zenodo.7181682.
Ordnance Survey Maps – Six-inch England and Wales, 1842–1952, National Library of Scotland: https://maps.nls.uk/os/6inch-england-and-wales/info1.html.
Ordnance Survey Maps – Six-inch 2nd and later editions, Scotland, 1892–1960, National Library of Scotland: https://maps.nls.uk/os/6inch-2nd-and-later/.
Ordnance Survey open roads data: https://www.ordnancesurvey.co.uk/business-government/products/open-map-roads.
Proceedings of the Belgian House of Representatives: https://plenum.be/.
Rail lines data collection (1851): M. Satchell, E. Wrigley, L. Shaw-Taylor, X. You and J. Martí-Henneberg, *1851 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852991, https://doi.org/10.5255/UKDA-SN-852991.
Rail lines data collection (1861): J. Martí-Henneberg, M. Satchell, X. You, L. Shaw-Taylor and E. Wrigley, *1861 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852992, https://doi.org/10.5255/UKDA-SN-852992.
Rail lines data collection (1881): J. Martí-Henneberg, M. Satchell, X. You, L. Shaw-Taylor and E. Wrigley, *1881 England, Wales and Scotland Rail Lines* [data collection]. UK Data Service, 2018. SN: 852993, https://doi.org/10.5255/UKDA-SN-852993.
Rumsey digital map collection: https://www.davidrumsey.com/.

Sanborn maps: https://www.loc.gov/collections/sanborn-maps/.
SoDuCo ('Social Dynamics in Urban Context: Open tools, models, and data – Paris and its suburbs, 1789–1950') datatsets: https://soduco.geohistoricaldata.org/catalog/.
Station to Station dataset: M. Coll Ardanuy et al., 'Station to Station'; https://rchs.org.uk/railway-passenger-stations-in-great-britain-a-chronology/, version 5.02 released September 2020 by the Railway and Canal Historical Society.
StopsGB dataset: Mariona Coll Ardanuy et al., 'StopsGB: Structured Timeline of Passenger Stations in Great Britain' (2021), https://doi.org/10.23636/wvva-3d67.

Code, tools and visualisations

AIKON computer vision platform: Albouy, Ségolene, Jade Norindr, Fouad Aouinti, Clara Grometto, Robin Champenois, Alexandre Guilbaud, Stavros Lazaris, Matthieu Husson and Mathieu Aubry, *AIKON: A Computer Vision Platform for Historians* (2024), https://github.com/Aikon-platform/aikon.
Allmaps mapping tool: https://allmaps.org/.
Animated map chronologies: https://observablehq.com/collection/@oliviafvane/map-animations-to-publish.
Beyond the Tracks code repository: https://github.com/Living-with-machines/beyondthetracks.
Book genre classifier: https://living-with-machines.github.io/genre-classification/intro.html.
Census geocoder: Joshua Rhodes, 'Geo-coding Addresses', CensusGeocoder pipeline code: https://github.com/Living-with-machines/CensusGeocoder.
Data analysis and visualisation of Zooniverse crowdsourcing tasks: King's Digital Lab, https://github.com/kingsdigitallab/lwm-davizct.
Image annotation tool: https://github.com/annotorious.
International Image Interoperability Framework: https://iiif.io/.
Language of Mechanisation data visualisation: Miguel Vieira, Tiffany Ong, Arianna Ciula, Barbara McGillivray, Mia Ridge and Nilo Pedrazzini, 'Language of Mechanisation', Observable, 4 March 2025, accessed 6 February 2026, https://observablehq.com/@jmiguelv/language-of-mechanisation.

Macromap: https://observablehq.com/@oliviafvane/macromap (data visualisation code).
Mapping tool: https://www.maptiler.com/nls/.
MapReader data tool guide: https://web.archive.org/web/20260219091958/https://mapreader.readthedocs.io/en/latest/index.html.
MapReader documentation: https://github.com/maps-as-data/MapReader/blob/main/.
MapReader 'railspace' ResNest model: https://huggingface.co/Livingwithmachines/mr_resnest101e_timm_pretrain.
Newspaper sampling tool: https://github.com/Living-with-machines/subsamplr.
Python record linking code: J. De Bruin, 'Python Record Linkage Toolkit: A Toolkit for Record Linkage and Duplicate Detection in Python. v0.14', Zenodo, 1 December 2019, https://doi.org/10.5281/zenodo.3559043.
SoDuCo: https://github.com/soduco (code repository).
Structural Topic Model tool: https://www.structuraltopicmodel.com/.
Tesseract: https://github.com/tesseract-ocr/tesseract (open-source OCR tool).

Other web resources

Deezer user forum: https://en.deezercommunity.com/ideas/who-is-the-composer-improved-labelling-of-music-551.
Design council guide: https://www.designcouncil.org.uk/our-work/skills-learning/the-double-diamond/.
Endings Project: https://endings.uvic.ca/index.html (project website).
Geopeuple: https://geopeuple.ign.fr/index.html (project website).
I-CeM documentation: https://www.campop.geog.cam.ac.uk/research/projects/icem/.
I-CeM documentation: Harry Smith and Robert J. Bennett, 'Research Data Supporting WP 6: Urban-Rural Classification Using Census Data, 1851–1911', Apollo – University of Cambridge Repository, 2019, https://doi.org/10.17863/CAM.44400.
I-CeM documentation: Kevin Schurer, 'I-CeM Lookup Table – Occupation Code'. Apollo – University of Cambridge Repository, 2024, https://doi.org/10.17863/CAM.106514.

Image tool users' forum: https://iiif.io/community/groups/maps/.
LaDéHiS: http://ladehis.ehess.fr/index.php?611 (project website).
Machines Reading Maps annotation guidelines: Valeria Vitale, 'Map Text Annotation Guidelines', https://github.com/machines-reading-maps/Tutorials-Newsletters/wiki/Map-Text-Annotation-Guidelines.
Miro software: https://miro.com.
Newspaper dataset guide: https://www.digitisednewspapers.net/histories/bl/ (accessed 4 April 2025).
Newspaper dataset guide: Mia Ridge and Nilo Pedrazzini, *Public Domain Newspaper Titles in Living with Machines* (Zenodo, 2024), https://doi.org/10.5281/zenodo.18480866.
Open Maps data forum: https://openmapsmeeting.nl/.
Vision of Britain: https://www.visionofbritain.org.uk/.
World Historical Gazetteer: https://whgazetteer.org/.
Zotero, v. 9.0.3, Corporation for Digital Scholarship, released 2026.

Primary

Anderson, Michael, *The 1851 Census: A Guide to the National Sample of the Enumerators' Returns* (Cambridge University Press, 1987).
Bell, Lady Florence, *At the Works: A Study of a Manufacturing Town* (T. Nelson & Sons, 1907).
Blake, William. *Milton: A Poem* (1804–8). https://blakearchive.org/copy/milton.d?descId=milton.d.illbk.01.
Booth, Charles, 'Condition and Occupations of the People of East London and Hackney, 1887', *Journal of the Royal Statistical Society* 51, no. 2 (1888): 276–331.
Booth, Charles, *Labour and Life of the People of London*, vol. 1, *East London* (Macmillan, 1889).
Booth, Charles, *Life and Labour of the People in London*, 3rd ser. (Macmillan, 1902).
Briavoinne, Natalis, *De l'industrie en Belgique: Sa Situation Actuelle. Causes de Décadence et de Prospérité* (Dubois, 1839).
Capitaine, Louis, César-François Cassini and the Cassini family, *Tableau de la Carte Générale de la France* (L. Capitaine, 1797) in César-François Cassini and the Cassini family, *Carte de France* (L. Capitaine, 1750–1815).
Carlyle, Thomas, 'Signs of the Times', *Edinburgh Review* (1829).

Census of England and Wales, 1891, General Report, vol. 4, 48. Accessed 2 March 2026, https://www.visionofbritain.org.uk/census/EW1891GEN/7.

Census of England and Wales, 1911, vol. 10, pt 1, *Occupations and Industries* [Cd. 7018] (HMSO, London, 1914), lxxvi–lxxvii.

Cobb, M.H., *The Railways of Great Britain: A Historical Atlas at a Scale of 1 inch to 1 mile* (Ian Allan Publishing, 2006).

Deacon's Newspaper Handbook and Advertisers' Guide, 1877–1904 (Charles Willam Deacon & Co.).

Dickens, Charles, *Hard Times* (Penguin, [1854] 1985).

Disraeli, Benjamin, *Sybil, or the Two Nations* (Henry Colman, 1845).

Engels, Friedrich, *The Condition of the Working Class in England in 1844* (Allen & Unwin, [1845] 1892).

Forster, E.M., *Howard's End* (Penguin, [1910] 1989).

Gaskell, Elizabeth, *Mary Barton* (Penguin, [1848] 2003).

Geddes, Patrick, *Cities in Evolution: An Introduction to the Town Planning Movement and to the Study of Civics* (Williams and Norgate, 1915).

Gissing, George, *Demos: A Story of English Socialism* (Smith Elder, 1886).

Gissing, George, *The Nether World* (Penguin, [1889] 1973).

Gissing, George, *New Grub Street* (Penguin, [1891] 1985).

Grossmith, George and Weedon, *The Diary of a Nobody* (Penguin, [1892] 1985).

Handy Newspaper List, 1895–1915 (Charles & Edwin Layton).

Joyce, James, *Ulysses* (Penguin, [1922] 2000).

LSE Library, Notebooks of Charles Booth's Survey of London, c.1886–1903. Accessed 2 March 2026, https://archives.lse.ac.uk/records/BOOTH/B.

May's British and Irish Press Guide, 1874–1889 (Frederick L. May & Co), later *Willing's*.

Mill, John Stuart, *Principles of Political Economy*, ed. by W.J. (Longmans, 1909).

Mitchell's Newspaper Press Directory, 1846–47, 1851, 1854–1976 (C. Mitchell) (later *The Newspaper Press Directory*).

Nicholson, J. Shield, *The Effects of Machinery on Wages* (Swan Sonnenschein, 1892).

Ordnance Survey of Great Britain, William Mudge, Thomas Colby and Benjamin Baker, *Index to the Ordnance Survey of England and Wales, Published in Sheets on a Scale of one inch to a Mile* (Ordnance Survey,

1870). Accessed 2 March 2026, https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~244652~5513977.

Oxford English Dictionary online (Oxford University Press, n.d.)

Rostow, Walt Whitman, *The Stages of Economic Growth: A Non-Communist Manifesto* (Cambridge University Press, 1960).

Rowntree, B. Seebohm, *Poverty: A Study of Town Life* (Macmillan, 1901).

Schumpeter, Joseph, *Business Cycles*, vol. 1 (Transaction Publishers, [1939] 1989).

Sell's Dictionary of the World Press, 1883–1921 (Sell's Advertising Agency).

Thomas, Edward, *In Pursuit of Spring* (Little Toller Books, [1914] 2016).

Toynbee, Arnold, *Lectures on the Industrial Revolution in England* (Rivingtons, 1884).

Wheeler, Rob, 'Introduction: A Six-Inch Cartobibliography'. Accessed 2 March 2026, https://www.charlesclosesociety.org/provo.

Willing's British and Irish Press Guide, from 1890 (James Willing), previously *May's Guide*.

Wordsworth, William, 'Outrage Done to Nature', in his poem *The Excursion*, Book 8, (Longman, 1814).

Secondary

Abernethy, Simon T., 'Opening up the Suburbs: Workmen's Trains in London 1860–1914', *Urban History*, 42, no. 1 (2015): 70–88, https://www.jstor.org/stable/26398343.

Abramitzky, R., R. Mill and S. Pérez, 'Linking Individuals across Historical Sources: A Fully Automated Approach', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, no. 2 (2019): 94–111.

Agarwal, Dhruv, Mor Naaman and Aditya Vashistha, 'AI Suggestions Homogenize Writing toward Western Styles and Diminish Cultural Nuances', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025): 1–21.

Ahnert, Ruth, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009175548.

Ahnert, Ruth, Katherine McDonough and Daniel C.S. Wilson, 'Show Me the Data: New Practices for Historical Sources', *Transactions of the*

Royal Historical Society (2026): 1–18. https://doi.org/10.1017/S0080440126100632.

Akgün, Ö., A. Dearle, G. Kirby, E. Garrett, T. Dalton, P. Christen, C. Dibben and L. Williamson, 'Linking Scottish Vital Event Records Using Family Groups', *Historical Methods: A Journal of Quantitative and Interdisciplinary* 53, no. 2 (2019): 130–46.

Alexander, David, 'Mapping the Past against the Present', *ACID (Archaeology and Conservation in Derbyshire)* 23 (January 2026): 10–11. Accessed 2 March 2026, https://www.peakdistrict.gov.uk/__data/assets/pdf_file/0023/210578/ACID_Newsletter_2026.pdf.

Algee-Hewitt, Mark, 'Computing Criticism: Humanities Concepts and Digital Methods', in *Computational Humanities*, ed. Lauren Tilton, David Mimno and Jessica Marie Johnson (University of Minnesota Press, 2024), https://dhdebates.gccuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/138dbec6-12b4-4728-9ca9-6540ad095530.

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti and Hannah Walser, 'Canon/Archive: Large-Scale Dynamics in the Literary Field', Literary Lab Pamphlet 11 (2016), https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet11.pdf.

Allen, R.C., *The British Industrial Revolution in Global Perspective* (Cambridge University Press, 2009).

Anon. [Olivia Vane] 'How Intrepid Victorian Surveyors Mapped the Length and Breadth of Britain', *The Economist*, 6 April 2023.

Anon., 'A History of Wednesfield, 4: The Nineteenth Century'. Accessed 2 March 2026, www.historywebsite.co.uk/articles/Wednesfield/nineteenth.htm.

Ardanuy, Mariona Coll, Kaspar Beelen, Jon Lawrence, Katherine McDonough, Federico Nanni, Joshua Rhodes, Giorgia Tolfo and Daniel C.S. Wilson, 'Station to Station: Linking and Enriching Historical British Railway Data', *Computational Humanities Research Workshop Proceedings* (2021): 249–65, https://ceur-ws.org/Vol-2989/long_paper29.pdf.

Ardanuy, Mariona Coll, Federico Nanni, Kaspar Beelen and Luke Hare, 'The Past Is a Foreign Place: Improving Toponym Linking for Historical Newspapers', *Proceedings of the Computational Humanities Research Conference 2023* (2023): 368–90. Accessed on 2 March 2026, http://ceur-ws.org/Vol-3558/.

Armaselu, Florentina, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė Oleškevičienė and Marieke van Erp , 'LL(O)D and NLP Perspectives on Semantic Change for Humanities Research', *Semantic Web* 13, no. 6 (2022).

Arnold, Taylor, and Lauren Tilton, 'Distant Viewing: Analyzing Large Visual Corpora', *Digital Scholarship in the Humanities* 34, supp. 1 (2019): i3–16, https://doi.org/10.1093/llc/fqz013.

Arnold, Taylor, and Lauren Tilton, *Distant Viewing* (MIT Press, 2023).

Ashton, T. S, *The Industrial Revolution* (Oxford University Press, 1948).

Atack, Jeremy, Fred Bateman and Mary Eschelbach Gregson, 'Matchmaker, Matchmaker, Make Me a Match', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 25, no. 2 (1992): 53–65.

Bailey, M., P.Z. Lin, A.R.S. Mohammed, P. Mohnen, J. Murray, M. Zhang and A. Prettyman, 'The Creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database Project', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56, no. 3 (2023): 138–59.

Beckert, Sven, *Empire of Cotton: A Global History* (Alfred A. Knopf, 2014).

Beelen, Kaspar and Natasha Kitcher, 'Heritage Weaver: Classifying, Searching, and Linking Museum Data with Multimodal AI', *Anthology of Computers and the Humanities* 3 (2025): 280–96.

Beelen, Kaspar, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan', *Digital Scholarship in the Humanities* 38, no. 1 (2023): 1–22, https://doi.org/10.1093/llc/fqac037.

Beelen, Kaspar, Jon Lawrence, Katherine McDonough and Daniel C.S. Wilson, 'Whose News? Critical Methods for Assessing Bias in Large Historical Datasets', *Computational Humanities Research* 1 (2025), https://doi.org/10.1017/chr.2025.10007.

Beelen, Kaspar and Daniel van Strien, 'Metadata Might Make Language Models Better', *arXiv* (2022): arXiv:2211.10086, https://doi.org/10.48550/arXiv.2211.10086.

Beelen, Kaspar, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo and Barbara McGillivray, 'When Time Makes Sense: A Historically-Aware Approach to Targeted Sense Disambiguation', in

Findings of the Association for Computational Linguistics: ACL-IJCNLP (2021): 2751–61.

Bell, Eamonn, Karina Rodriguez Echavarria and Jeyan Thiyagalingam, 'CCP-AHC Roadmap Open Draft', Zenodo, 11 September 2025, https://doi.org/10.5281/zenodo.17099176.

Berg, Maxine, 'What Difference Did Women's Work Make to the Industrial Revolution?' *History Workshop Journal* 35, no. 1 (1993): 22–44.

Berg, Maxine, *Luxury and Pleasure in Eighteenth-Century Britain* (Oxford University Press, 2005).

Berg, Maxine, and Pat Hudson, 'Rehabilitating the Industrial Revolution', *Economic History Review* 45, no.1 (1992): 24–50.

Bergel, Giles, Pip Willcox, Guyda Armstrong, James Baker, Arianna Ciula, Nicholas Cole, Julianne Nyhan, Mia Ridge, Oscar Selp, Claire Taylor, Pip Thornton, Elizabeth Williamson and Jane Winters, 'Sustaining Digital Humanities in the UK', Zenodo, 25 September 2020, https://doi.org/10.5281/zenodo.4046267.

Bergs, Alexander, and Laurel J. Brinton, ed., *English Historical Linguistics* (De Gruyter Mouton, 2012).

Blaxill, Luke, 'Why Do Historians Ignore Digital Analysis? Bring on the Luddites', *Political Quarterly* 94, no. 2 (2023): 279–89.

Blevins, Cameron, 'Digital History's Perpetual Future Tense', in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016), https://doi.org/10.5749/9781452963761.

Bode, Katherine, 'The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History', *Modern Language Quarterly* 78, no. 1 (2017): 77–106, https://doi.org/10.1215/00267929-3699787.

Bode, Katherine, 'The Archive', in *The Cambridge Companion to Literature in a Digital Age*, ed. Adam Hammond (Cambridge University Press, 2024): 89–106, https://doi.org/10.1017/9781009349567.007.

Bode, Katherine, 'Why You Can't Model Away Bias', *Modern Language Quarterly* 81, no. 1 (2020): 95–124.

Bogart, Dan, 'The Transport Revolution in Industrializing Britain', in *The Cambridge Economic History of Modern Britain: Volume I, 1700–1870*, ed. R. Floud, J. Humphries and P. Johnson (Cambridge University Press, 2014).

Bogart, Dan, Xuesheng You, Eduard J. Alvarez-Palau, Max Satchell and Leigh Shaw-Taylor, 'Railways, Divergence, and Structural Change in 19th-century England and Wales', *Journal of Urban Economics* 128 (2022): 1–23, https://doi.org/10.1016/j.jue.2021.103390.

Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky and Percy S. Liang, 'Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?', *Advances in Neural Information Processing Systems* 35 (2022): 3663–78.

Bonato, Nicolò, 'LwM Digital Residency: Visualising the Press Directories Dataset', September 2023. Accessed 2 March 2026, https://livingwithmachines.ac.uk/lwm-digital-residency-visualising-the-press-directories-dataset/.

Brake, Laurel, 'Nineteenth-Century Newspaper Press Directories: The National Gallery of the British Press', *Victorian Period Review* 48 (2015): 569–90.

Brewer, Charlotte, 'The Electronification of the Oxford English Dictionary', *Dictionaries: Journal of the Dictionary Society of North America* 25 (2004): 1–43.

Bronstein, Jamie L., *The Happiness of the British Working Class* (Stanford University Press, 2023).

Buckles, Kasey, Adrian Haws, Joseph Price and Haley E.B. Wilbert, 'Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project', National Bureau of Economic Research Working Paper Series, 2023, Working Paper 31671.

Cannadine, David, 'Residential Differentiation in Nineteenth-Century Towns', in *The Structure of Nineteenth Century Cities*, ed. James Henry Johnson and Colin G. Pooley (Croom Helm, 1982), 235–51.

Carus-Wilson, E.M., 'An Industrial Revolution of the Thirteenth Century', *Economic History Review* 11 (1941): 39–60.

Caviglia, Giorgio and Nicole Coleman, 'Idiographic Network Visualizations', *Leonardo* 49, no. 5 (2016): 447.

Chaloner, William Henry, 'Was There a Decline of the Industrial Spirit in Britain 1850–1939?', *Transactions of the Newcomen Society* 55 (1985 for 1983–84): 214.

Chiang, Yao-Yi, and Craig A. Knoblock, 'Recognizing Text in Raster Maps', *Geoinformatica* 19, no. 1 (2015): 1–27, https://doi.org/10.1007/s10707-014-0203-9.

Chiang, Yao-Yi, Weiwei Duan, Stefan Leyk, Johannes H. Uhl and Craig A. Knoblock, *Using Historical Maps in Scientific Studies: Applications, Challenges, and Best Practices* (Springer, 2020), https://doi.org/10.1007/978-3-319-66908-3.

Christopher, A.J., 'The Quest for a Census of the British Empire, c. 1840–1940', *Journal of Historical Geography* 34, no. 2 (2008): 268–85.

Clapham, John H., *An Economic History of Modern Britain: The Early Railway Age, 1820–1850* (Cambridge University Press, 1926).

Clifford, Jim, Beatrice Alex, Colin M. Coates, Ewan Klein and Andrew Watson, 'Geoparsing History: Locating Commodities in Ten Million Pages of Nineteenth-Century Sources', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49, no. 3 (2016): 115–31, https://doi.org/10.1080/01615440.2015.1116419.

Coleman, Niamh NicGhabhann, Zenobie Garrett and Frances Kane, '"That's a Powerful Map": Shared Authority, Public Engagement, and the Archives of the First Ordnance Survey of Ireland', *Journal of Historical Geography* 88 (June 2025): 65–73, https://doi.org/10.1016/j.jhg.2024.12.006.

Coll Ardanuy, Mariona, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel C.S. Wilson and Barbara McGillivray, 'Living Machines: A Study of Atypical Animacy', *Proceedings of the 28th International Conference on Computational Linguistics* (2020): 4534–45.

Corcoran, Evangeline, Kasra Hosseini, Laura Siles, Smita Kurup and Sebastian Ahnert, 'Automated Dynamic Phenotyping of Whole Oilseed Rape (*Brassica napus*) Plants from Images Collected under Controlled Conditions', *Frontiers in Plant Science* 16 (2025): 1443882, https://doi.org/10.3389/fpls.2025.1443882.

Corcoran, Evangeline, Laura Siles, Smita Kurup and Sebastian Ahnert, 'Automated Extraction of Pod Phenotype Data from Micro-Computed Tomography', *Frontiers in Plant Science* 14 (2023): 1120182.

Cordell, Ryan, David A. Smith, Abby Mullen, Jonathan D. Fitzgerald and Avery Blankenship, *Going the Rounds: Virality in Nineteenth-Century American Newspapers* (University of Minnesota Press, Manifold Platform, 2025), https://manifold.umn.edu/projects/going-the-rounds.

Couldry, Nick, and Ulises A. Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating it for Capitalism* (Stanford University Press, 2020).

Cowie, Jefferson, *Stayin' Alive: The 1970s and the Last Days of the Working Class* (The New Press, 2010).

Crafts, N.F.R., *British Economic Growth during the Industrial Revolution* (Clarendon Press, 1985).

Crawford, Kate, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021).

Cunningham, Niall, and Ian Gregory, 'Hard to Miss, Easy to Blame? Peacelines, Interfaces and Political Deaths in Belfast During the Troubles', *Political Geography* 40 (2014): 64–78, https://doi.org/10.1016/j.polgeo.2014.02.004.

Curtis, Bruce. *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875* (University of Toronto Press, 2002).

Daunton, M.J., *Progress and Poverty: An Economic and Social History of Britain 1700–1850* (Oxford University Press, 1995).

Daunton, Martin J., *State and Market in Victorian Britain: War, Welfare and Capitalism* (Boydell Press, 2008).

Daunton, Martin J., *Wealth and Welfare: An Economic and Social History of Britain, 1851–1951* (Oxford University Press, 2007).

Davies, Alun C., *The Rise and Decline of England's Watchmaking Industry, 1550–1930* (Routledge, 2022).

Deane, Phyllis, *The First Industrial Revolution* (Cambridge University Press, 1967).

Deane, Phyllis, and W.A. Cole, *British Economic Growth, 1688–1959: Trends and Structure* (Cambridge University Press, 1962).

Dennis, Richard, *English Industrial Cities of the Nineteenth Century: A Social Geography* (Cambridge University Press, 1984).

de Vries, Jan, *The Industrious Revolution: Consumer Behaviour and the Household Economy* (Cambridge University Press, 2008).

D'Ignazio, Catherine, and Lauren F. Klein, *Data Feminism* (MIT Press, 2023).

Dobson, James, and Scott Sanders, 'Distant Approaches to the Printed Page', *Digital Studies / Le Champ Numérique* 12, no. 1 (2022), https://doi.org/10.16995/dscn.8107.

Dombrowski, Quinn, Tassie Gniady and David Kloster, 'Introduction to Jupyter Notebooks', *The Programming Historian* (2019; updated 2025), https://doi.org/10.46430/phen0087.

Dombrowski, Quinn, Tassie Gniady, David Kloster, Megan Meredith-Lobay, Jeffrey Tharsen and Lee Zickel, in 'Voices from the Server Room: Humanists in High-Performance Computing', in *Computational Humanities*, ed. Lauren Tilton, David Mimno and Jessica Marie Johnson (Minnesota University Press, 2024), 247–64. Accessed 2 March 2026, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/57bf0197-fd4a-44ce-a542-78883631cacb.

Drucker, Johanna, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly* 5, no. 1 (2011), https://dhq-static.digitalhumanities.org/pdf/000091.pdf.

Drucker, Johanna, 'Sustainability and Complexity: Knowledge and Authority in the Digital Humanities', *Digital Scholarship in the Humanities* 36, supp. 2 (2021): ii86–94, https://doi.org/10.1093/llc/fqab025.

Dyche, Jill, 'Big Data "Eurekas!" Don't Just Happen', *Harvard Business Review* Blog 20 (November 2012).

Dyos, H.J., *Victorian Suburb: A Study of the Growth of Camberwell* (Leicester University Press, 1973).

Dyos, H.J. and D.A. Reeder, 'Slums and Suburbs', in *The Victorian City*, ed. H.J. Dyos and M. Wolff (Routledge and Kegan Paul, 1973), 359–86.

Edney, Matthew H., 'Academic Cartography, Internal Map History, and the Critical Study of Mapping Processes', *Imago Mundi* 66, supp. 1 (2014): 83–106, https://doi.org/10.1080/03085694.2014.947853.

Edney, Matthew H., *Cartography: The Ideal and Its History* (University of Chicago Press, 2019).

Edney, Matthew H., 'JB Harley (1932–1991): Questioning Maps, Questioning Cartography, Questioning Cartographers', *Cartography and Geographic Information Systems* 19, no. 3 (1992): 175–8, https://doi.org/10.1559/152304092783762290.

Eliot, Simon, 'Circulating Libraries in the Victorian Age and After', in *The Cambridge History of Libraries in Britain and Ireland*, vol. 3, *1850–2000*, ed. Alistair Black and Peter Hoare (Cambridge University Press, 2006), 125–46.

Evans, A.K.B., and J.V. Gough, eds, *The Impact of the Railway on Society in Britain: Essays in Honour of Jack Simmons* (Ashgate, 2003).

Farrell, Henry, Alison Gopnik, Cosma Shalizi and James Evans, 'Large AI Models Are Cultural and Social Technologies', *Science* 387, no. 6739 (2025): 1153–6 (p. 1156).

Ferrie, J.P., 'A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29, no. 4 (1996): 141–56.

Finlay, Ronald, and Kevin H. O'Rourke, *Power and Plenty: Trade, War, and the World Economy in the Second Millennium* (Princeton University Press, 2008).

Fittschen, Elisabeth, Sabrina Li, Tom Lippincott, Leshem Choshen and Craig Messner, 'Pretraining Language Models for Diachronic Linguistic Change Discovery', *arXiv* (2025): arXiv:2504.05523, https://doi.org/10.48550/arXiv.2504.05523.

Fleet, Chris, 'Creating, Managing, and Maximising the Potential of Large Online Georeferenced Map Layers', *e-Perimetron* 14, no. 3 (2019): 140–49. Accessed 2 March 2026, https://www.e-perimetron.org/Vol_14_3/Fleet.pdf.

Fleet, Chris, 'Life Beyond the Paper Graphic Index: Evaluating New Graphic Retrieval Technologies for the Future Map Library', *Liber Quarterly* 18, no. 2 (2008): 255–70. Accessed 2 March 2026, https://liberquarterly.eu/article/view/10507/11163.

Fleet, Chris, 'Understanding User Needs: A Case Study from the National Library of Scotland', Digital Preservation Coalition Technology Watch Guidance Note (2022). Accessed 2 March 2026, https://www.dpconline.org/docs/technology-watch-reports/2774-understanding-user-needs-national-library-of-scotland/file.

Fox, Alan, 'Industrial Relations in Nineteenth-Century Birmingham', *Oxford Economic Papers* 7, no. 1 (1955): 57–70.

Franke, Max, John Markus, Mortiz Knabben, Jana Keck, Tanja Blascheck and Steffen Koch, 'LilyPads: Exploring the Spatiotemporal Dissemination of Historical Newspaper Articles', *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*

(VISIGRAPP 2020), vol. 3 (IVAPP): 17–28, https://doi.org/10.5220/0008871400170028.

Fyfe, Paul, 'An Archaeology of Victorian Newspapers', *Victorian Periodicals Review* 49, no. 4 (2016): 546–77.

Gagnier, Regenia, *Subjectivities: A History of Self-Representation in Britain, 1832–1920* (Oxford University Press, 1991).

Galison, Peter, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997).

Galli, S., D. Theodoridis and K. Rönnbäck, 'Reconstructing a Slave Society: Building the DWI Panel, 1760–1914', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 57, no. 3 (2024): 163–84.

Gershuny, J.I., and R.E. Pahl, 'Britain in the Decade of the Three Economies', *New Society*, 3 January 1980.

Gershuny, J.I., and R.E. Pahl, 'Work Outside Employment', *New Universities Quarterly* 34, no. 1 (1979): 120–35.

Gieryn, Thomas F., 'City as Truth-Spot: Laboratories and Field-Sites in Urban Studies', *Social Studies of Science* 36, no. 1 (2006): 5–38, https://doi.org/10.1177/0306312705054526.

Glover, Frederick J., 'The Rise of the Heavy Woollen Trade of the West Riding of Yorkshire in the 19th Century', *Business History* 4, no. 1 (1961): 1–21.

Goeken, Ron, Lap Huynh, T.A. Lynch and Rebecca Vick, 'New Methods of Census Record Linking', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44, no. 1 (2011): 7–14.

Görlach, M., *English in Nineteenth-Century England* (Cambridge University Press, 1999).

Griffin, Emma, *Bread Winner: An Intimate History of the Victorian Economy* (Yale University Press, 2020).

Griffin, Emma, *Liberty's Dawn: A People's History of the Industrial Revolution* (Yale University Press, 2013).

Griffin, Emma, 'The Making of the Chartists: Popular Politics and Working-class Autobiography in Early Victorian Britain', *English Historical Review* 129, no. 538 (2014), 578–605, https://doi-org.libproxy.york.ac.uk/10.1093/ehr/ceu078.

Groot, Jerome De, 'The Genealogy Boom: Inheritance, Family History, and the Popular Historical Imagination', in *The Impact of History? Histories at the Beginning of the Twenty-First Century*, ed. Pedro Ramos Pinto and Bertrand Taithe (Routledge, 2015), 21–33.

Guldi, Jo, 'Critical Search: A Procedure for Guide Reading in Large-Scale Textual Corpora', *Journal of Cultural Analytics* 3, no. 1 (2018): 1–34, https://doi.org/10.22148/16.030.

Guldi, Jo, *The Dangerous Art of Text Mining: A Methodology for Digital History* (Cambridge University Press, 2023), https://doi.org/10.1017/9781009263016.

Guo, Yanzhu, Guokan Shang, Michalis Vazirgiannis and Chloé Clavel, 'The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text', in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024): 3589–604.

Hamilton, William L., Jure Leskovec and Dan Jurafsky, 'Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2016), 1489–501.

Hammond, J.L., and Barbara Hammond, *The Skilled Labourer* (Longmans, Green & Co., 1919).

Haraway, Donna, 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies* 14, no. 3 (1988): 575–99, https://doi.org/10.2307/3178066.

Harley, J.B., *The New Nature of Maps: Essays in the History of Cartography*, ed. Paul Laxton (Johns Hopkins University Press, 2001).

Harley, J.B., and David Woodward, eds, *The History of Cartography*, 6 vols (University of Chicago Press, 1987–2019).

Harrison, A.E., 'The Origins and Growth of the UK Cycle Industry to 1900', *The Journal of Transport History* 6, no. 1 (1985): 43–5.

Hawkins, Angus, *British Party Politics, 1852–1886* (Macmillan, 1998).

Hawkins, Angus, '"Parliamentary Government" and Victorian Political Parties, c.1830–c.1880', *English Historical Review* 104 (1989): 638–69.

Heblich, Stephan, Stephen J. Redding and Daniel M. Sturm, 'The Making of the Modern Metropolis: Evidence from London', *Quarterly Journal of Economics* 135, no. 4 (2020): 2059–133, https://doi.org/10.1093/qje/qjaa014.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles and Catherine A. Fitch, 'A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55, no. 1 (2022): 12–29.

Hellyer, Roger, and Richard Oliver, *The First Ordnance Survey Map: The One-Inch Old Series of England and Wales* (Charles Close Society, 2015).

Hennock, E.P., 'Concepts of Poverty in the British Social Surveys from Charles Booth to Arthur Bowley', in *The Social Survey in Historical Perspective, 1880–1940*, ed. Martin Bulmer, Kevin Bales and Kathryn Kish Sklar (Cambridge University Press, 1991), 189–216.

Herlihy, David V., *Bicycle: The History* (Yale University Press, 2004).

Hewitt, Martin, *The Dawn of the Cheap Press in Victorian Britain: The End of the 'Taxes on Knowledge', 1849–1869* (Bloomsbury, 2014).

Hewitt, Rachel, *Map of a Nation: A Biography of the Ordnance Survey* (Granta, 2010).

Higgs, Edward, 'The Rise of the Information State: The Development of Central State Surveillance of the Citizen in England, 1500–2000', *Journal of the History of Sociology* 14, no. 2 (2001): 175–97.

Hills, Thomas T., Eugenio Proto and Daniel Sgroi, 'Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books', *Nature Human Behaviour* 3, no. 12 (2019): 1271–5.

Hitchcock, Tim, 'Confronting the Digital: Or How Academic History Writing Lost the Plot', *Cultural and Social History* 10 (2013): 9–23, https://doi.org/10.2752/147800413X13515292098070.

Hobbs, Andrew, *A Fleet Street in Every Town: The Provincial Press in England, 1855–1900* (Open Book Publishers, 2018), https://doi.org/10.11647/OBP.0152.

Hobson, J.A., 'The Influence of Machinery upon Employment', *Political Science Quarterly* 8, no. 1 (1893): 97–123.

Horrell, Sara, and Jane Humphries, 'Old Questions, New Data, and Alternative Perspectives: Families' Living Standards in the Industrial Revolution', *Journal of Economic History* 52, no. 4 (1992): 849–80.

Hosseini, Kasra, Kaspar Beelen, Giovanni Colavizza and Mariona Coll Ardanuy, 'Neural Language Models for Nineteenth-century English', *Journal of Open Humanities Data* 7 (2021): 22.

Hosseini, Kasra, Katherine McDonough, Daniel van Strien, Olivia Vane and Daniel C.S. Wilson, 'Maps of a Nation? The Digitized Ordnance Survey for New Historical Research', *Journal of Victorian Culture* 26, no. 2 (April 2021): 284–99, https://doi.org/10.1093/jvcult/vcab009.

Hosseini, Kasra, Daniel C.S. Wilson, Kaspar Beelen and Katherine McDonough, 'MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale', in *Proceedings of the*

6th ACM SIGSPATIAL International Workshop on Geospatial Humanities (Association for Computing Machinery, 2022): 8–19, https://doi.org/10.1145/3557919.3565812.

Hou, Liwen and David A. Smith, 'Detecting Syntactic Change with Pre-Trained Transformer Models', in *Findings of the Association for Computational Linguistics: EMNLP 2023* (2023), pp. 3564–74.

Hudson, Pat, ed., *Regions and Industries: A Perspective on the Industrial Revolution in Britain* (Cambridge University Press, 1989).

Impett, Leonardo, and Fabian Offert, 'There Is a Digital Art History', *Visual Resources* 38, no. 2 (2022): 186–209, https://doi.org/10.1080/01973762.2024.2362466.

Inikori, Joseph E., *Africans and the Industrial Revolution in England: A Study in International Trade and Economic Development* (Cambridge University Press, 2002).

Jackson, Alan, *Semi-Detached London: Suburban Development, Life and Transport, 1900–39* (Allen & Unwin, 1973).

Jaro, Matthew A., 'Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida', *Journal of the American Statistical Association* 84, no. 406 (1989): 414–20, https://doi.org/10.1080/01621459.1989.10478785.

Johnson, James, and Colin G. Pooley, 'The Internal Structure of the Nineteenth-Century British City: An Overview', in *The Structure of Nineteenth Century Cities*, ed. James Johnson and Colin G. Pooley (Croom Helm, 1982).

Jones, Eric, *The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia*, 3rd ed. (Cambridge University Press, 2003).

Jones, Gareth Stedman, *Outcast London: A Study in the Relationship between Classes in Victorian Society* (Clarendon Press, 1971).

Joyce, Patrick, *The State of Freedom: A Social History of the British State since 1800* (Cambridge University Press, 2013).

Kadish, Alon, 'Arnold Toynbee (1852–1883)', *Oxford Dictionary of National Biography*, (Oxford University Press, 2004).

Kaplan, Frédéric, and Isabella de Lenardo, 'The Advent of the 4D Mirror World', *Urban Planning* 5, no. 2 (2020): 307–10, https://doi.org/10.17645/up.v5i2.3133

Kaplan, Frédéric, and Isabella de Lenardo, 'Building a Mirror World for Venice', in *The Aura in the Age of Digital Materiality: Rethinking*

Preservation in the Shadow of an Uncertain Future (Factum Foundation, 2020), 197–201.
Kay, C., and K.L. Allan, *English Historical Semantics* (Edinburgh University Press, 2015).
Kee, Kevin, and Timothy Compeau, eds, *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History* (University of Michigan Press, 2019), https://doi.org/10.3998/mpub.9964786.
Kellett, John R., *The Impact of Railways on Victorian Cities* (Routledge and Kegan Paul, 1969).
Kemman, Max, *Trading Zones of Digital History* (De Gruyter, 2021).
Kenny, S., 'Sub-regional Specialization in the Lancashire Cotton Industry, 1884–1914: A Study in Organizational and Locational Change', *Journal of Historical Geography* 8, no. 1 (1982): 41–63.
Kent, Alexander James, Soetkin Vervust, Imre Josef Demhardt and Nick Millea, eds, *Mapping Empires: Colonial Cartographies of Land and Sea: 7th International Symposium of the ICA Commission on the History of Cartography, 2018* (Springer International Publishing, 2020).
Ketabgian, Tamara, *The Lives of Machines: The Industrial Imaginary in Victorian Literature and Culture* (University of Michigan Press, 2011).
Killick, Roberta, Paul Fearnhead and Idris A. Eckley, 'Optimal Detection of Changepoints with a Linear Computational Cost', *Journal of the American Statistical Association* 107, no. 500 (2012): 1590–98.
Kimberley, Damien, *Coventry's Bicycle Heritage* (The History Press, 2015).
King, Edmund, 'Digitisation of Newspapers at the British Library', *Serials Librarian* 49 (2005): 165–81, https://doi.org/10.1300/J123v49n01_07.
Kitchin, Rob, 'Big Data and Human Geography: Opportunities, Challenges and Risks', *Dialogues in Human Geography* 3, no. 3 (2013): 262–7.
Kitchin, Rob, 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society* 1, no. 1 (2014): 1–12, https://doi.org/10.1177/2053951714528481.
Klein, Lauren, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton and David Mimno, 'Provocations from the Humanities for Generative AI Research', *arXiv* (2025), https://doi.org/10.48550/arXiv.2502.19190.

Kleinberg, Jon and Manish Raghavan, 'Algorithmic Monoculture and Social Welfare', *Proceedings of the National Academy of Sciences* 118, no. 22 (2021): e2018340118.

Knowles, Anne Kelly, 'Emerging Trends in Historical GIS', *Historical Geography* 33 (2005): 7–13.

Knowles, Anne Kelly, 'Historical Geographic Information Systems and Social Science History', *Social Science History* 40, no. 4 (2016): 741–50, https://doi.org/10.1017/ssh.2016.29.

Knowles, Anne Kelly, ed., *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship* (Environmental Systems Research Institute, 2008).

Krause, Monika, *Model Cases: On Canonical Research Objects and Sites* (University of Chicago Press, 2021).

Lan, T., and P. Longley, 'Geo-Referencing and Mapping 1901 Census Addresses for England and Wales', *ISPRS International Journal of Geo-Information* 8, no. 8 (2019), https://doi.org/10.3390/ijgi8080320.

Lan, T., and P. Longley, 'Urban Morphology and Residential Differentiation across Great Britain, 1881–1901', *Annals of the American Association of Geographers* 111, no. 6 (2021): 1796–1815, https://doi.org/10.1080/24694452.2020.1859982.

Lancaster, Bill, 'Who's a Real Coventry Kid? Migration into Twentieth Century Coventry' in *Life and Labour in a 20th Century City: The Experience of Coventry*, ed. Bill Lancaster and Tony Mason (Cryfield Press, n.d. [c.1980]), 57–80.

Lancaster, Bill, and Tony Mason, eds, *Life and Labour in a 20th Century City: The Experience of Coventry* (Cryfield Press, n.d. [c.1980]).

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper team and Nello Cristianini, 'Content Analysis of 150 Years of British Periodicals', *Proceedings of the National Academy of Sciences* 114, no. 4 (2017): e457–65.

Lavin, Matthew, 'Why Digital Humanists Should Emphasize Situated Data over Capta', in *The Companion to Digital Humanities in Practice*, ed. Constance Crompton, Laura Estill, Richard J. Lane, Ray Siemens (Routledge, 2025), 121–36, https://doi.org/10.4324/9781003327677.

Lee, Alan J., *The Origins of the Popular Press, 1855–1914* (Croom Helm, 1976).

Leslie, Sarah J., Sangeet Khemlan, Sandeep Prasada and Sam Glucksberg, 'Conceptual and Linguistic Distinctions Between

Singular and Plural Generics', in *Proceedings of the 31st Annual Cognitive Science Society* (2009): 479–84.

Li, Zekun, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu and Nathalie Abadie, 'ICDAR 2024 Competition on Historical Map Text Detection, Recognition, and Linking', in *Document Analysis and Recognition – ICDAR 2024: 18th International Conference Proceedings, Athens, Greece, Aug 30–Sept 4, 2024* (Springer Verlag, 2024), 363–80, https://doi.org/10.1007/978-3-031-70552-6_22.

Lin, Yijun, and Yao-Yi Chiang, 'Hyper-Local Deformable Transformers for Text Spotting on Historical Maps', in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2024), 5387–97, https://doi.org/10.1145/3637528.3671589.

Lin, Yijun, Solenn Tual, Zekun Li, Leeje Jang, Yao-Yi Chiang, Jerod Weinman, Joseph Chazalon et al., 'ICDAR 2025 Competition on Historical Map Text Detection, Recognition, and Linking', in *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025* (Springer Verlag, 2025), 568–85, https://doi.org/10.1007/978-3-032-04630-7_33.

Lincoln, Matthew, Julia Corrin, Emily Davis and Scott B. Weingart, 'CAMPI: Computer-Aided Metadata Generation for Photo Archives Initiative' (2020), Carnegie Mellon University Libraries Whitepaper, https://doi.org/10.1184/R1/12791807.

Linkon, Sherry Lee, *The Half-Life of Deindustrialization: Working-Class Writing about Economic Restructuring* (University of Michigan Press, 2018).

Malm, Andreas, *Fossil Capital: The Rise of Steam Power and the Roots of Global Warming* (Verso Books, 2016).

Manlove, Colin, 'Charles Kingsley, H.G. Wells, and the Machine in Victorian Fiction', *Nineteenth-Century Literature* 48, no. 2 (1993): 212–39.

Mari, Alda, Claire Beyssade and Fabio Del Prete, eds, *Genericity* (Oxford University Press, 2012).

Martí-Henneberg, Jordi, 'Geographical Information Systems and the Study of History', *The Journal of Interdisciplinary History* 42, no. 1 (2011): 1–13, https://doi.org/10.1162/JINH_a_00202.

Massey, Catherine G., Katie R. Genadek, J. Trent Alexander, Todd K. Gardner and Amy O'Hara, 'Linking the 1940 U.S. Census with

Modern Data', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51, no. 4 (2018): 246–57.

McDonough, Katherine, 'Maps as Data', in *Computational Humanities*, ed. Lauren Tilton, David Mimno and Jessica Marie Johnson (Minnesota University Press, 2024), 99–126, https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/992f5fc5-9a0d-4e9a-863a-adc1de9a678e.

McDonough, Katherine, Kaspar Beelen, Daniel C.S. Wilson and Rosie Wood, 'Reading Maps at a Distance: Texts on Maps as New Historical Data', *Imago Mundi* 76, no. 2 (2024): 296–307, https://doi.org/10.1080/03085694.2024.2453336.

McDonough, Katherine, Ludovic Moncla and Matje van de Camp, 'Named Entity Recognition Goes to Old Regime France', *International Journal of Geographical Information Science* 33, no. 12 (2019): 2498–522, https://doi.org/10.1080/13658816.2019.1620235.

McDonough, Katherine, and Valeria Vitale, 'Introduction', in 'Maps and Machines: Recent Perspectives on Humanities Research Using AI and Maps', ed. Katherine McDonough and Valeria Vitale, Forum, *Imago Mundi* 76, no. 2 (2024): 282, https://doi.org/10.1080/03085694.2025.2453329.

McGillivray, Barbara, *How to Use Word Embeddings for Natural Language Processing* (SAGE Research Methods: Doing Research Online, 2022), https://doi.org/10.4135/9781529609578.

McGrory, David, *A Century of Coventry: Events, People and Places Over the 20th Century* (Sutton Publishing, 2007).

Meijers, Martijn, and Jules Schoonman, 'Mapping the Edge: A Novel Approach to Georeferencing Historical Map Series', *e-Perimetron* 20, no. 1 (2025): 1–51, http://www.e-perimetron.org/Vol_20_1/Meijers_et_al.pdf.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Liebermann Aiden, 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science* 331, no. 6014 (2011): 176–82.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean, 'Efficient Estimation of Word Representations in Vector Space', *arXiv* (2013), https://doi.org/10.48550/arXiv.1301.3781.

Milligan, Ian, *History in the Age of Abundance: How the Web Is Transforming Historical Research* (McGill-Queen's University Press, 2019).
Milligan, Ian, *The Transformation of Historical Research in the Digital Age* (Cambridge University Press, 2022).
Millward, Robert, *Private and Public Enterprise in Europe* (Cambridge University Press, 2005).
Mitch, David F., *The Rise of Popular Literacy in Victorian England: The Influence of Private Choice and Public Policy* (University of Pennsylvania Press, 1991).
Mitchell, B.R., *British Historical Statistics* (Cambridge University Press, 1988).
Mokyr, Joel, 'Is There Still Life in the Pessimist Case? Consumption during the Industrial Revolution', *Journal of Economic History* 48, no. 1 (1988): 69–92.
Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn, 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis* 16, no. 4 (2008): 372–403.
Muehrcke, Phillip C., *Map Use: Reading Analysis and Interpretation*, 2nd ed. (JP Publications, 1986), vii.
The Multigraph Collective, *Interacting with Print: Elements of Reading in the Era of Print Saturation* (University of Chicago Press, 2018).
Murrieta-Flores, Patricia, and Ian N. Gregory, 'Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology', *Open Archaeology* 1, no. 1 (2015): 166–75, https://doi.org/10.1515/opar-2015-0010.
Nef, J.U., 'The Progress of Technology and the Growth of Large-Scale Industry in Great Britain, 1540–1640', *Economic History Review* 5, no. 1 (1934): 3–24.
Nef, J.U., *The Rise of the British Coal Industry*, vol. 1. (George Routledge & Sons, 1932).
Nelson, Robert K., 'Mapping Environmental Inequalities: Using MapReader to Uncover Mid-Century Industrial Burdens', *Imago Mundi* 76, no. 2 (2024): 284–90, https://doi.org/10.1080/03085694.2024.2453333.
Norcliffe, Glen, 'G-COT: The Geographical Construction of Technology', *Science, Technology, & Human Values* 34 no. 4 (2009): 461–5.
Norcliffe, Glen, *Ride to Modernity: The Bicycle in Canada, 1869–1900* (University of Toronto Press, 2001), 47–58.

O'Brien, Patrick, 'Path Dependency, or Why Britain Became an Industrialized and Urbanized Economy Long before France', *Economic History Review* 49, no. 2 (1996): 213–49.

Oliver, Richard, *The Ordnance Survey in the Nineteenth Century: Maps, Money, and the Growth of Government* (Charles Close Society, 2014).

O'Malley, Tom, 'Mitchell's "Newspaper Press Directory" and the Late Victorian and Early Twentieth-Century Press', *Victorian Periodicals Review* 48, no. 4 (2015): 591–606.

Pechenick, E.A., C.M. Danforth and P.S. Dodds, 'Characterizing the Google Books Corpus', *PLoS ONE* 10, no. 10 (2015). https://doi.org/10.1371/journal.pone.0137041.

Pedrazzini, Nilo, and Barbara McGillivray, 'Machines in the Media: Semantic Change in the Lexicon of Mechanization in 19th-century British Newspapers', in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities* (Association for Computational Linguistics, 2022), 85–95.

Pelletier, Monique, *Les Cartes des Cassini: La Science au Service de l'État et des Provinces*, new ed. (Éditions du CTHS, 2013).

Periti, Francesco and Stefano Montanelli, 'Lexical Semantic Change through Large Language Models: A Survey', *ACM Computing Surveys* 56, no. 11 (2024): 1–38.

Petitpierre, Remi, and Paul Guhennec, 'Effective Annotation for the Automatic Vectorization of Cadastral Maps', *Digital Scholarship in the Humanities* 38, no. 3 (2023): 1227–37, https://doi.org/10.1093/llc/fqad006.

Petitpierre, Rémi, Frédéric Kaplan and Isabella di Lenardo, 'Generic Semantic Segmentation of Historical Maps', *CHR 2021: Computational Humanities Research Conference Proceedings* (17 November 2021): 228–48, https://ceur-ws.org/Vol-2989/long_paper27.pdf.

Piper, Andrew, Chad Wellmon and Mohamed Cheriet, 'The Page Image: Towards a Visual History of Digital Documents', *Book History* 23, no. 1 (2020): 365–97, https://doi.org/10.1353/bh.2020.0010.

Poovey, Mary, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (University of Chicago Press, 1998).

Porter, Catherine, Paul Atkinson and Ian N. Gregory, 'Space and Time in 100 Million Words: Health and Disease in a Nineteenth-Century

Newspaper', *International Journal of Humanities and Arts Computing* 12, no. 2 (2018): 196–216, https://doi.org/10.3366/ijhac.2018.0222.

Posner, Miriam, 'What's Next: The Radical, Unrealized Potential of Digital Humanities', in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/a22aca14-0eb0-4cc6-a622-6fee9428a357.

Putnam, Lara, 'The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast', *The American Historical Review* 121, no. 2 (2016): 377–402.

Ramsay, Stephen, *Reading Machines: Toward an Algorithmic Criticism* (University of Illinois Press, 2011), 10, https://doi.org/10.5406/illinois/9780252036415.001.0001.

Raphael, Lutz, *Beyond Coal and Steel: A Social History of Western Europe after the Boom*, trans. Kate Tranter (Polity, 2023).

Reeder, David A., *Charles Booth's Descriptive Map of London Poverty 1889* (London Topographical Society, 1984).

Řehůřek, Radim, and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (European Language Resources Association, 2010), 45–50.

Rhodes, J., 'Geo-coding Addresses in Historic British Census Data: An Open Methodology', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 58, no. 1 (2025): 31–53, https://doi.org/10.1080/01615440.2024.2431491.

Richardson, Kenneth, *Twentieth-Century Coventry* (Macmillan, 1972).

Ridge, Mia, 'From Prams to Parliament: What Was a Machine? Help Us Find Out', *Living with Machines*, 4 December 2020, https://livingwithmachines.ac.uk/from-prams-to-parliament-what-was-a-machine-help-us-find-out/.

Ridge, Mia, 'From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing', *Curator: The Museum Journal* 56, no. 4 (2013): 435–50, https://doi.org/10.1111/cura.12046.

Ridge, Mia, 'Outreach and Marketing for Crowdsourcing Tasks', *Living with Machines*, 27 June 2024, https://livingwithmachines.ac.uk/outreach-and-marketing-for-crowdsourcing-tasks/.

Ridge, Mia, 'What Is a "Machine" Anyway? Help Us Describe Them', *Living with Machines*, 27 October 2020, https://livingwithmachines.ac.uk/what-is-a-machine-anyway-help-us-find-out/.

Ridge, Mia, Nilo Pedrazzini, Miguel Vieira, Arianna Ciula and Barbara McGillivray, 'Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project', *Journal of Open Humanities Data* 10, no. 1 (2024): 1–9, https://doi-org.libproxy.york.ac.uk/10.5334/johd.195.

Rieger, B., *Technology and the Culture of Modernity in Britain and Germany, 1890–1945* (Cambridge University Press, 2005).

Ringler, Hannah, 'Computation and Hermeneutics: Why We Still Need Interpretation to Be by (Computational) Humanists', *Computational Humanities* 6 (2024), https://dhdebates.gc.cuny.edu/read/computational-humanities-5c64bbab-d7ca-41be-8f87-f26117a9a20f/section/cdccb7af-e7cd-4225-932a-f12da4214b1a.

Risam, Roopika, and Alex Gil, 'Introduction: The Questions of Minimal Computing', *Digital Humanities Quarterly* 16, no. 2 (2022), https://dhq.digitalhumanities.org/vol/16/2/000646/000646.html.

Roberts, Robert, *The Classic Slum: Salford Life in the First Quarter of the Century* (Manchester University Press, 1971).

Rockwell, Geoffrey, and Stéfan Sinclair, *Hermeneutica: Computer-Assisted Interpretation in the Humanities* (Cambridge, MA: MIT Press, 2016), https://doi.org/10.7551/mitpress/9522.001.0001.

Roland, Edwin, Richard Jean So and Hoyt Long, 'The Social AI Author: Modeling Creativity and Distinction in Simulated Cultural Fields', *AI & Society* (2025), https://doi.org/10.1007/s00146-025-02790-0.

Rosen, Paul, *Framing Production: Technology, Culture and Change in the British Bicycle Industry* (MIT Press, 2002).

Rosenberg, Nathan, 'Technological Change in the Machine Tool Industry, 1840–1910', *The Journal of Economic History* 23, no. 4 (1963): 414–43.

Ruggles, Steven, 'Linking Historical Censuses: A New Approach', *History and Computing* 14, nos. 1–2 (2002): 213–24.

Sakoe, H., and S. Chiba, 'Dynamic Programming Algorithm Optimization for Spoken Word Recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, no. 1 (1978): 43–9, https://doi.org/10.1109/TASSP.1978.1163055.

Samuel, Raphael, 'Workshop of the World: Steam Power and Hand Technology in Mid-Victorian Britain', *History Workshop Journal* 3 (1977): 6–72.

Savage, Michael, *The Dynamics of Working-Class Politics: The Labour Movement in Preston, 1880–1940* (Cambridge University Press, 1987).

Schivelbusch, Wolfgang, *The Railway Journey: The Industrialization of Time and Space* (Berg, 1986).

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky and Nina Tahmasebi, 'SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection', *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020* (2020): 1–23.

Schürer, Kevin, and Joe Day, 'Migration to London and the Development of the North–South Divide, 1851–1911', *Social History* 44, no. 1 (2019): 26–56.

Schürer, Kevin, Eilidh M. Garrett, Hannaliis Jaadla, and Alice Reid, 'Household and Family Structure in England and Wales, 1851–1911', *Continuity and Change* 33, no. 3 (2018): 365–411.

Schwartz, Robert, Ian Gregory and Thomas Thevenin, 'Spatial History: Railways, Uneven Development, and Population Change in France and Great Britain, 1850–1914', *Journal of Interdisciplinary History* 42 (2011): 53–88, https://doi.org/10.1162/JINH_a_00205.

Seymour, W.A., *A History of the Ordnance Survey* (Dawson, 1980).

Shaw-Taylor, Leigh, 'Diverse Experiences: The Geography of Adult Female Employment in England and the 1851 Census', in *Women's Work in Industrial England*, ed. Nigel Goose (Local Population Studies Society, 2007), 29–50.

Shaw-Taylor, Leigh, and E.A. Wrigley, 'Occupational Structure and Population Change', in *The Cambridge Economic History of Modern Britain*, vol. 1, ed. Roderick Floud, Jane Humphries and Paul A. Johnson (Cambridge University Press, 2014), 53–88.

Smith, Harry, and Robert J. Bennett, 'Urban-Rural Classification Using Census Data, 1851–1911', Drivers of Entrepreneurship and Small Business, Working Paper 6 in Working paper series from ESRC project ES/M010953 (2017), https://doi.org/10.17863/CAM.15763.

Smith, H., R. Bennett and D. Radicic, 'Towns in Victorian England and Wales: A New Classification', *Urban History* 45, no. 4 (2018): 568–94, https://doi.org/10.1017/S0963926818000020.

Smits, Thomas, and Melvin Wevers, 'A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections', *Digital*

Scholarship in the Humanities 38, no. 3 (2023): 1267–80, https://doi.org/10.1093/llc/fqad008.

Snow, C.P., *The Two Cultures*, with an introduction by Stefan Collini (Cambridge University Press, 1998).

Snydman, Stuart, Robert Sanderson and Tom Cramer, 'The International Image Interoperability Framework (IIIF): A Community and Technology Approach for Web-based Images', in *Archiving Conference* 12 (Society for Imaging Science and Technology, 2015): 16–21.

Spufford, Francis, and Jenny Uglow, *Cultural Babbage: Technology, Time and Invention* (Faber and Faber, 1996).

Stapel, Rombert, and Ivo Zandhuis, 'Linked Data for Modelling and Replicating the Knowledge Production Process in Data-driven Humanities Research', *Digital Scholarship in the Humanities* 40, supp. 1 (2025): i100–i107, https://doi.org/10.1093/llc/fqae038.

Strien, Daniel van, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray and Giovanni Colavizza, 'Assessing the Impact of OCR Quality on Downstream NLP Tasks', in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence* 1: ARTIDIGH (2020): 484–96, https://doi.org/10.5220/0009169004840496.

Sussman, Herbert, *Victorians and the Machine: The Literary Response to Technology* (Harvard University Press, 1968).

Sussman, Herbert, *Victorian Technology: Invention, Innovation, and the Rise of the Machine* (Praeger, 2009).

Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu and Simon Hengchen, eds, *Computational Approaches to Semantic Change* (Language Science Press, 2021).

Tang, Xuri, 'A State-of-the-Art of Semantic Change Computation', *Natural Language Engineering* 24, no. 5 (2018): 649–76.

Taylor, Joanna E., and Ian N. Gregory, *Deep Mapping the Literary Lake District: A Geographical Text Analysis* (Lewisburg: Bucknell University Press, 2022), https://doi.org/10.2307/j.ctv2v55bsf.

Thompson, E.P., *The Making of the English Working Class* (Penguin, [1963] 1976).

Thoms, D.W., and T. Donnelly, 'Coventry's Industrial Economy, 1880–1980', in *Life and Labour in a Twentieth-Century City: The Experience of Coventry*, ed. Bill Lancaster and Tony Mason (Cryfield Press [c.1980]), 12–21.

Tolfo, Giorgio, Olivia Vane, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, David Beavan and Katherine McDonough, 'Hunting for Treasure: Living with Machines and the British Library Newspaper Collection', in *Digitised Newspapers: A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*, ed. Estelle Bunout, Maud Ehrmann and Frédéric Clavert (De Gruyter Oldenbourg, 2022), 23–46.

Tolfo, Giorgia, Olivia Vane, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, David Beavan and Katherine McDonough, 'The Minimum Research Outcome: A Mechanism for Generating and Managing Projects in Labs', in *Digital Humanities and Laboratories: Perspectives on Knowledge, Infrastructure and Culture*, ed. Urszula Pawlicka-Deger and Christopher Thomson (Routledge, 2023), 193–208.

Trouillot, Michel-Rolph, *Silencing the Past: Power and the Production of History* (Beacon Press, 1995).

Turing Way Community, *The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research*, Version 1.2.3 (2025), https://doi.org/10.5281/zenodo.15213042.

Turnock, David, *An Historical Geography of Railways in Great Britain and Ireland* (Ashgate, 1998).

Uhl, Johannes H., Stefan Leyk, Yao-Yi Chiang, Weiwei Duan and Craig A. Knoblock, 'Automated Extraction of Human Settlement Patterns from Historical Topographic Map Series using Weakly Supervised Convolutional neural networks', *IEEE Access* 8 (2019): 6978–96, https://doi.org/10.1109/ACCESS.2019.2963213.

Uhl, Johannes H., Stefan Leyk, Yao-Yi Chiang and Craig A. Knoblock, 'Towards the Automated Large-Scale Reconstruction of Past Road Networks from Historical Maps', *Computers, Environment and Urban Systems* 94 (2022): 101794, https://doi.org/10.1016/j.compenvurbsys.2022.101794.

Umoja Noble, Safiya, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018).

Underwood, Ted, 'Dear Humanists: Fear Not the Digital Revolution: Advances in Computing Will Benefit Traditional Scholarship', *Chronicle of Higher Education*, 27 March 2019.

Underwood, Ted, *Distant Horizons: Digital Evidence and Literary Change* (University of Chicago Press, 2019).

Vaienti, Beatrice, Rémi Petitpierre, Isabelle di Lenardo and Frédéric Kaplan, 'Machine-Learning-Enhanced Procedural Modeling for 4D

Historical Cities Reconstruction', *Remote Sensing* 15, no. 13 (2023): 3352, https://doi.org/10.3390/rs15133352.

Vaienti, Beatrice, Isabelle di Lenardo and Frédéric Kaplan, 'Exploring Cartographic Genealogies through Deformation Analysis: Case Studies on Ancient Maps and Synthetic Data', *Cartography and Geographic Information Science* 52, no. 5 (2025): 557–77, https://doi.org/10.1080/15230406.2024.2424891.

Vane, Olivia, 'Finding Words in Maps Part 2: Seeing the Results', *Living with Machines*, 22 August 2019, https://livingwithmachines.ac.uk/finding-words-in-maps-part-2-seeing-the-results/; archived at https://web.archive.org/web/20260223125954/https://livingwithmachines.ac.uk/finding-words-in-maps-part-2-seeing-the-results/.

Vane, Olivia, 'Macromap: Interactive Maps in Time', 16 November 2021, https://livingwithmachines.ac.uk/macromap-interactive-maps-in-time/.

Vernon, James, *Distant Strangers: How Britain Became Modern* (University of California Press, 2014).

Victoria County History, *A History of the County of Warwick*, vol. 8, *The City of Coventry and Borough of Warwick*, ed. W.B. Stephens (Oxford University Press for the Institute of Historical Research, 1969), *British History Online,* https://www.british-history.ac.uk/vch/warks/vol8/pp162-189.

Vincent, David, *Bread, Knowledge and Freedom: A Study of Nineteenth-Century Working-Class Autobiography* (Europa Publications, 1981).

Vincent, David, *Literacy and Popular Culture: England, 1750–1914* (Cambridge University Press, 1989).

Vitale, Valeria, Pau de Soto, Rainer Simon, Elton Barker, Leif Isaksen and Rebecca Kahn, 'Pelagios – Connecting Histories of Place. Part I: Methods and Tools', *International Journal of Humanities and Arts Computing* 15, nos. 1–2 (2021): 5–32, https://doi.org/10.3366/ijhac.2021.0260.

Webb, Sidney, and Beatrice Webb, *Industrial Democracy*, 2 vols (Longmans, 1898).

Wernimont, Jacqueline, *Numbered Lives: Life and Death in Quantum Media* (MIT Press, 2019).

West Midland Group, *Conurbation: A Planning Survey of Birmingham and the Black Country* (Architectural Press, 1948).

Westerling, Kalle, Kaspar Beelen, Tim Hobson, Katherine McDonough, Nilo Pedrazzini, Daniel C.S. Wilson and Ruth Ahnert, 'LwMDB: Open Metadata for Digitised Historical Newspapers from British Library Collections', *Journal of Open Humanities Data* 11, no. 1 (2025), article 32. https://doi.org/10.5334/johd.304.

White, Jerry, *The Worst Street in North London: Campbell Bunk, Islington, between the Wars* (Routledge & Kegan Paul, 1986).

Wiener, Martin J., *English Culture and the Decline of the Industrial Spirit, 1850–1980* (Cambridge University Press, 1981).

Willenhall History Society, 'Willenhall: The Town of Locks and Keys', www.willenhallhistory.co.uk/society/locksandkeys.htm.

Williams, Chris A., 'Risk on the Roads: Police, Motor Traffic and the Management of Space, c. 1900–1950', in *Governing Risks in Modern Britain: Danger, Safety and Accidents, c. 1800–2000*, ed. Tom Crook and Mike Esbester (Palgrave Macmillan, 2016), 195–219.

Williams, Eric, *Capitalism and the Slave Trade* (University of North Carolina Press, 1944).

Williamson, J.G., 'Was the Industrial Revolution Worth It? Disamenities and Death in Nineteenth-Century British Towns', *Explorations in Economic History* 19, no. 3 (1982): 221–45.

Wilson, Daniel C.S., 'Finding Words in Maps', *Living with Machines*, 30 July 2019, https://livingwithmachines.ac.uk/finding-words-in-maps/; archived at https://web.archive.org/web/20260223130245/https://livingwithmachines.ac.uk/finding-words-in-maps/.

Wilson, Daniel C.S., 'Working at Scale: What Do Computational Methods Mean for Research Using Cases, Models and Collections?' *Science Museum Group Journal* 18 (2023), https://doi.org/10.15180/221805.

Wilson, Daniel C.S., Mariona Coll Ardanuy, Kaspar Beelen, Barbara McGillivray and Ruth Ahnert, 'The Living Machine: A Computational Approach to the Nineteenth-Century Language of Technology', *Technology and Culture* 64, no. 3 (2023): 875–902.

Windelband, Wilhelm, 'History and Natural Science', *Theory & Psychology* 8, no. 1 (1998): 5–22, https://doi.org/10.1177/0959354398081001.

Winkler, William E., 'String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage', in *American Statistical Association, Proceedings of the Section on Survey Research Methods* (1990): 354–59.

Wissen, Leon van, Manjusha Kuruppath and Lodewijk Petram, 'Unlocking the Research Potential of Early Modern Dutch Maps', *European Journal of Geography* 16, no. 1 (2025): s12–17, https://doi.org/10.48088/ejg.si.spat.hum.l.wis.12.17.

Wood, Denis, with John Fels and John Krygier, *Rethinking the Power of Maps* (The Guilford Press, 2010).

Wood, Rosie, Kasra Hosseini, Kalle Westerling, Andrew Smith, Kaspar Beelen, Daniel C.S. Wilson and Katherine McDonough, 'MapReader: Open Software for the Visual Analysis of Maps', *Journal of Open Source Software* 9, no. 101 (2024): 6434, https://doi.org/10.21105/joss.06434.

Woodward, David, ed., *Five Centuries of Map Printing* (University of Chicago Press, 1975).

Wrigley, E.A., *Continuity, Chance and Change: The Character of the Industrial Revolution in England* (Cambridge University Press, 1988).

Wrigley, E.A., *The Path to Sustained Growth: England's Transition from an Organic Economy to an Industrial Revolution* (Cambridge University Press, 2016).

You, X., D. Bogart, E. Alvarez, A.E.M. Satchell and L. Shaw-Taylor, 'Transport Development and Urban Population Change in the Age of Steam: A Market Access Approach', *Campop Transport Working Papers* 9, at https://www.campop.geog.cam.ac.uk/research/occupations/outputs/preliminary/marketaccessandsteam.pdf.

Xing, Jia, Wei Ren, Du Lei, Lin Zhao, Xue Qin, Hongdang Shao, Wenjie Li, Yirui Han, Zike Yu, Zheng Xu, Rui Yin, Jiantao Yuan, Jun Wang, Wei Chen, Jun Xu, Xiaoping Zhou, Cheng Yang, Wei Zhou and Binbin Zhou, 'Multimodal AI for Yuan Buddhist Sculpture Chronology and Style', *npj Heritage Science* 13, no. 1 (2025): 443.

Zhu, Z., 'Like Father Like Son? Intergenerational Immobility in England, 1851–1911', *The Journal of Economic History* 84, no. 4 (2024): 1143–73.

Unpublished secondary

Abernethy, Simon, 'Class, Gender and Commuting in Greater London, 1880–1940' (Unpublished PhD thesis, University of Cambridge, 2015).

Harrison, Anthony Edward, 'Growth, Entrepreneurship and Capital Formation in the United Kingdom's Cycle and Related Industries, 1870–1914' (Unpublished PhD thesis, University of York, 1977).

Lenardo, Isabella di, and Frédéric Kaplan, 'Venice Time Machine: Recreating the Density of the Past', paper presented at the annual ADHO Digital Humanities Conference, Sydney, 2015.

Lilley, Kirsty, Katherine McDonough, David Alexander and Rosie Wood, 'Landscape Change and Conservation with MapReader in England's National Parks', poster presented at the CoSeC (Computational Science Centre for Research Communities) Annual Conference, Manchester, 3 December 2025.

McDonough, Katherine, Kaspar Beelen and Daniel C.S. Wilson, 'Neighbourhood Walks: A New Semantic Topology for Historical Map Text', paper presented at Computational Humanities Research 2025 (Lightning Talk Session), Luxembourg, 11 December 2025.

McDonough, Katherine, Bert Spaan, Kaspar Beelen, Jules Schoonman, Valeria Vitale, Daniel C.S. Wilson and Rosie Wood, 'Text on Maps: From Collection Discovery to Research Data', paper presented at Fantastic Futures 2025, London, UK, 4 December 2025, https://doi.org/10.23636/54k0-ny43.

Oliveira, Sofia Ares, Isabelle di Lenardo, Bastien Tourenc and Frédéric Kaplan, 'A Deep Learning Approach to Cadastral Computing', paper presented at the annual ADHO Digital Humanities Conference, Utrecht, 2019.

Ridge, Mia, Nilo Pedrazzini and Zooniverse volunteers, 'What "Machine" Words Changed When?' *Living with Machines Talk*, January 2023, https://www.zooniverse.org/projects/bldigital/living-with-machines/talk/2795/2773552?comment=4555192&page=1.

Taylor, Sally-Ann, 'Tradition and Change: the Sheffield Cutlery Trades, 1870–1914' (Unpublished PhD thesis, University of Sheffield, 1988).

Thomas, R., 'The Victorian Intergenerational Migrant Cohort: Socioeconomic Outcomes, 1851–1901' (Unpublished MPhil thesis, University of Oxford, 2018).

Index

accuracy rates, 146, 194, 243
AddressGB, 138, 140, 147
address linking to 'real-world' locations, 140
administrative units on census, 140
advertisers, 41, 43
aggregation of data, 132, 133, 139, 140
agriculture, 9
AI (Artificial Intelligence)
  Alan Turing Institute, 21
  bias, 81
  black box techniques, 242
  ChatGPT, 1, 305, 311
  ethics, 308
  'fairer,' 61
  generative AI, 306–7, 310
  historians and the future of, 305–18
  humanities-AI interface, 307
  MapReader, 118
  multimodality, 313
  music streaming, 33
  source criticism at scale, 38
  technological innovation for research, 1–2, 3, 4
  training data, 307
AIKON, 87n74
Alan Turing Institute, 21, 22, 107, 108
Alexander, David, 116
Algee-Hewitt, Mark, xxiv, 83n15, 112, 118
algorithmic criticism, 111–12
algorithmic transparency, 242
algorithms
  algorithmic criticism, 111–12
  algorithmic transparency, 242
  automated labelling, 78
  Fightin' Words algorithm, 58
  semantic change detection research, 6, 264–5, 287, 288
Allmaps, 109, 116
amenity and disamenity
  balance of interests, 137–8
  Camberwell, 163
  environmental disamenities, 136, 138, 148, 163, 165, 167, 168
  finely calibrated decision-making, 155–6
  hazardous sites, 161
  local versus national, 132
  Middlesbrough, 165
  mutual reinforcement of, 163
  occupational data, 153, 154
  social class, 138, 150
  type of railspace, 168
  Victorian rail infrastructure, 135–7, 148, 152
American English, 270, 278–80, 281, 282, 287, 288
analogue search, 34
analysis, emphasis on, 22–3
ancestral occupations, 220
animated map chronologies, 69–72, 101–2
annotation
  annotation tasks for volunteers, 274–6
  human annotation, 74, 77–8, 258, 274–6, 287–8, 290–1
  inter-annotator agreement, 289, 290, 291, 292
  keyword-based annotation, 78
  manual annotation, 47, 105, 107, 117, 146, 286, 289–98
  MapReader, 100, 145
  maps, 94, 104–5, 107, 114
  metadata, 47, 76–9
  semantic annotation, 47, 257, 258, 287–8, 290–1
  team annotations, 114
anonymisation, 142
Antoniak, Maria, 306
ArcGIS, 93, 99, 110
Armitage, David, 12
Arts and Humanities Research Council (AHRC), 307
Ashton, T.S., 9–10
At the Works (Bell, 1907), 157
atypical animacy, 311
authoritative sources, 41
authorship model, 25, 315, 317
autobiography, 10, 11

automated detection of semantic change, 258, 259–74
automated record linkage, 185, 187
automatic classification tools, 94
automatic parsing, 44

backward linking, 48, 195–6
balance of interests, 137–8
'barrier effects,' 156, 167, 168
Beelen, Kaspar, 24, 34, 43, 44, 47, 116
Bell, Lady Florence, 157
BERT language models, 47, 78
bias
  articulation of, 57
  census, 144
  confirmation bias, 34
  'Environmental Scan' method, 17
  epistemological bias, 99
  hidden biases, 55
  index maps, 63
  maps, 63, 111, 144
  measuring, 55–8
  missing word prediction tasks, 311
  modelling alternate worldviews, 59–61
  newspapers, 35, 36, 309
  no analysis without provenance, 308–9
  recency bias, 13
  statistical bias, 56
  transparency, 80
  visibility of, 309
big data
  false sense of security, 56
  limitations of, 36
'bike' semantic change, 280, 281–5, 288, 291, 292
biology, 12, 107, 117
Birmingham, 180–97, 203, 207–19, 235, 238–9, 241
Blaxill, Luke, 12
Bloom, Leopold, 135, 150, 169
Bode, Katherine, 36, 112
Bogart, Dan, 140
Bonato, Nicolò, 54
book corpora, 73–9. *See also* Google Books; Microsoft Books
Booth, Charles, 133, 136, 160, 163, 168
Bridging Responsible AI Divides (BRAID) programme, 307
Britain as the first industrial nation, 8–9, 19
British Library
  and Alan Turing Institute, 21
  books, 74, 78
  British Library Labs, 55
  British Newspaper Archive (BNA), 13, 16, 37, 55, 257
  collaborative working, 315
  data services, 75
  digitised newspapers, 22, 37, 55, 260
  and FindMyPast, 56
  Imaging Studio, 43
  maps, 19, 72, 101
  Metadata Standards team, 77
British Nation, 56
British Newspaper Archive (BNA), 16, 37, 55, 257
Brock, André, 306
browsing, 34
business districts. *See* central business district workers

Camberwell, 156, 158, 161, 162, 168
Cambridge Group for the History of Population and Social Structure, 11
Campbell Road, Upper Holloway, 161
Cannadine, David, 134
car industry, 183, 185, 197, 230–1, 232, 245
'car' semantic change, 273, 274, 281–5, 286, 288, 291, 292
carto-bibliographies, 64, 73
cartography
  Environmental Scan, 36
  history of, 110, 144
  MapReader, 92, 99, 101, 108
  multimodal cartography, 94, 108
'Cassini' map, 64
cataloguing
  books, 78
  maps, 61, 65–6, 73, 107
census
  administrative units on census, 140
  bias, 144
  Census Enumerator Books (CEBs), 18
  combining with map data, 65

cycle trade project, 184–97
as data source, 17–18, 20
digitisation, 4
geo-coded 1901 census data, 138, 139–43
individual-level data, 139, 140, 179, 184–97
industrialisation, 3
levels of analysis, 139
microcensus data, 113, 139, 140
occupational data, 153, 243
parish-level census data, 140, 142, 143, 158
railspace, 115, 147–56
record linkage, 185, 187–97, 241–2
social group-level census data, 179–80
street-level data, 11, 138–47, 156–67, 168, 313
surrogate markers for wealth/income, 148–9
US Census Tree, 187
Census Tree, 185, 187
central business district workers
Camberwell, 161, 162
Middlesbrough, 164–5
national trends, 168
Salford, 165–7
surrogate markers for wealth/income, 154, 155
Upper Holloway, 159, 160
chain identifiers, 48
changepoint detection, 264–76
ChatGPT, 1, 305, 311
Chatham News, 43
Chiang, Yao-Yi, 107
choices about where to live, 135, 148, 165
circulation information, 55
cities, 132–3, 135, 136, 144, 313
citizen science, 258, 275, 287–8
'civics,' 133
classification, 76–9, 102, 105
'Classifying 19th Century British Library Books using Crowdsourcing and Machine Learning,' 74
class-making, 10
cliometrics, 11, 12
close reading, 12, 13, 35, 116, 258, 267–71, 287, 289
cloud platforms, 67
clustering techniques, 168–9, 265, 268
'coach' semantic change, 270, 271, 280, 281–5, 287, 288, 291, 292
coal, 9
coal depots, 158, 161
coarse-grained data, 98, 145
code repositories, 23, 291–2, 317
Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project (Ahnert et al, 2023), 5, 314
collaborative working, 4–5, 20–6, 93, 109, 118, 257, 308, 313–16, 317. *See also* radical collaboration
Coll Ardanuy, Mariona, 24
collections
digital collection building, 57
maps, 65–6, 95, 144
Microsoft Books, 74
missing data, 36, 37, 39, 40, 56, 74–5, 101, 309
under-sampling in data collections, 56, 75
searching digital collections, 33–5, 38–9, 93
series maps as, 63–4, 65–6
colonialism, 9, 20
commercial partnerships, 4
community-sourced references, 147. *See also* crowdsourcing
commuters, 135, 148, 150, 152, 154, 157
composite maps, 67
composite metalworkers, 184–5, 196, 202–6, 213–19, 225–6, 235–7, 245–8
computational humanities, 5, 6, 11–15, 38, 118
computational linguistics, 21, 23, 59, 91, 258, 275, 311
computational literary studies, 92
computational map studies, 91–2
computer vision (CV), 21–4, 65, 93, 95, 98, 100, 145
'condition of England' debate, 79
confidence scores, 105, 278
confirmation bias, 34
conservative politics, 52, 56
Construction sector, 198
content analysis, 58–9
contextualisation of semantic shifts, 287

contextualised content analysis, 58–9
contextualised datasets
  articulation of bias, 57
  contextualisation at scale, 14
  Environmental Scan, 33–4, 35, 36, 38–9
  historical reference resources, 41
  large language models (LLMs), 310
  maps, 72–3, 111
  newspapers, 308
  situated knowledge, 310–11
  useability of data, 5
contextualised meanings of words, 258–9, 260
contextual sources, identification of, 40–3
contours and biases, measuring, 55–8
convergence, 132, 138–47
convergence experiments, 25, 147–56, 313
Corcoran, Evangeline, 107
corpus linguistics, 257, 288
cosine similarity scores, 265
Coventry, 180–97, 203, 207–29, 235, 238–9, 240, 241
Coventry Machinists Company, 182
Crafts, Nick, 8
crediting research labour, xvii–xx, 6, 25, 315
critical humanities, 96, 101, 111
critical search, 66, 67–9
critical techniques and tools, 14, 35, 36
cross-referenced data, 23
crowdsourcing
  book digitisation, 74, 77–8
  GLAM (gallery, library, archive and museum) organisations, 38
  semantic annotation, 257, 258, 274–6, 287–8, 290–8
cultural lag, 3, 4
Culturomics, 12, 73–4
curation practices, 58
curators, 5, 38, 40, 58, 65, 72–3, 79, 101, 257, 314

Data/Culture: Building Sustainable Communities around Arts and Humanities datasets and software (2023–25), 108, 317
data-driven history, 4, 11–15
data 'extraction,' 111
data granularity, 98
data interoperability, 47. *See also* International Image Interoperability Framework (IIIF)
Daunton, Martin, 137
David Rumsey collection, 103, 107
decontextualised information, 34, 37, 38–9
deep learning, 93
'deep scan', 39, 40, 54, 55, 61–79, 80, 309
democratisation of print culture, 13
Dennis, Richard, 134
Depot Road, Middlesbrough, 163
deprived areas, 160, 163, 168
derived data, 96–7, 140, 142
diachronic change, studying, 51
diachronic word embedding models, 258, 261–2, 286, 288
diatopic (geographic) lexical variation, 258, 272–4, 287, 289
digital accessibility, 5, 297
digital cartography, 110
digital collection building, 57
digital history, 4–5, 6, 11–15, 23
digital humanities (DH), 94, 95, 100, 118, 188, 275, 306
digital literary studies, 36
digital research infrastructure, 100
digitisation
  books, 12–13, 37, 74
  census, 140
  of gaps in collections, 101
  maps, 4, 64–6, 67
  missing data, 33–4
  newspapers, 11, 21
  priorities for, 57, 309
  transparency, 74
digitised newspapers
  British Library, 16–17, 22, 56
  British Newspaper Archive (BNA), 16, 37, 55, 257
  contextualised datasets, 35
  data-driven history, 4
  JISC corpus, 55–6
  linking to *Directories* records, 48
  paper and ink quality, 59
  primary markets, 57
  semantic change, 260

d'Ignazio, Catherine, 99
disadvantaged people, 148. *See also* marginalised voices; poverty
disciplinary silos, 4
discoverability, 33, 107, 317
distant reading, 35, 36, 59, 81
distilbert-base-cased language model, 78
divergence calculations, 56, 59
domain knowledge, 34
domestic industry, 9
Doré, Gustave, 136, 137
double diamond process, 295
drop-down capitals, 45
Drucker, Johanna, 99
Duckworth, George H., 163
Dynamic Time Warping, 265, 268
Dyos, H.J., 134, 136–7, 156, 161

economic growth, 7–8
edge cases, 104, 114
Edney, Matthew, 97
education, 256
employment and the rail infrastructure, 136, 137, 148, 152, 153, 154, 155
Engels, Friedrich, 7, 133
enhanced metadata
  contextualised datasets, 41, 43
  Environmental Scan, 44, 47, 54, 55, 58, 80, 309
  language of mechanisation, 258, 260, 272, 290
  maps, 69, 73
  modelling alternate worldviews, 59–61
environmental disamenities, 136, 138, 148, 163, 165, 167, 168
Environmental Scan method, 14, 17, 33–89, 272, 308
environmental scientists, 95
epistemological assumptions, 310
epistemological shifts, 92, 95–106, 110–16, 117
epistemological trading zones, 314
error post-correction, 262, 263–4, 289
error rates, 115, 243
ethnicity, 163, 168
evaluation accuracy scores, 78, 115, 194
Evans, James, 305–6

experimental approaches, 60, 94, 97, 101, 106–7, 113, 139, 305
exporting datasets, 105–6
external data linkages, 48, 73, 257, 265, 287
extracted data, 96

false negatives, 115, 194
false positives, 115, 142, 276
family historians, 57, 179
Farrell, Henry, 305–6
female workers, 143, 185, 194, 225, 227, 244
fiction books, 76–9
Fightin' Words algorithm, 58
FindMyPast, 13, 16, 17, 18, 56, 57
fine-tuning models, 44–7, 78, 100, 105, 115, 145, 146, 311, 312
flat data structures, 98
Fleet, Chris, 101
flooding, 165
fonts, 44, 45
Forster, E.M., 135, 150
'fourth industrial revolution, 2
France, 64
funding, 16, 18, 55, 65, 97, 107, 305, 307, 312, 316–17

Gagnier, Regenia, 10
Gale, 16, 17
Galison, Peter, 314
gazetteers, 66, 73
GB1900, 24, 142–3
'gear' semantic change, 267, 271, 287
Geddes, Patrick, 133, 134, 169
generalisability, 93–4, 242, 258, 316–18
genre, 76–9
gentrification, 180
geo-coded 1901 census data, 142
geo-coded passenger rail stations dataset, 147
geo-coordinates, 48
geo-data, 19, 24
Geographical Information Systems (GIS), 21, 23, 65, 67, 93, 94, 98, 99, 113, 114, 140
geographic information scientists, 95
geography, 110, 134
GeoHistoricalData, 64

geopandas, 113
GéoPeuple, 64
georeferencing information, 66, 92, 102, 109, 116, 139, 142
geospatial coordinates, 102, 105, 109
geospatial information overload, 91, 92
Gissing, George, 138, 155
GitHub, 291, 315
GLAM (gallery, library, archive and museum) organisations, 6, 38, 48, 57, 81
goods depots, 158, 161
Google Books, 12, 74
Gopnik, Alison, 305–6
Görlach, Manfred, 259
Gospel Oak, 158, 159
gradual industrialisation, 7–8
Grahame, Kenneth, 284
Griffin, Emma, 10–11
Grossmiths, George and Weedon, 156
'ground truths,' 96
Guldi, Jo, 12, 13, 67, 69
gun-making, 182, 185, 197, 220, 222

harm adjudication, 137–8
heavy industry, 153, 154, 157, 167
Heritage Made Digital (HMD), 16, 56
hidden biases, 55
Higgs, Edward, 195–6, 199–201, 204–6, 207–12, 214–19, 221, 224, 227, 230, 236–9
Hillman, William, 180
Hillman & Herbert, 182
historians' ability to handle data, 4
Historical GIS, 95–6, 97
historical linguistics, 257, 259, 270, 290
historical reference metadata, 40
historical reference resources, 39, 40, 41, 43–7, 310
historical sociology, 12
Hobbs, Andrew, 57
Hobson, Tim, 60
Hollington Street, 161, 163
Holmes-Wong, Deborah, 107
Hosseini, Kasra, 24
hotels and inns, 152, 163
house-level data, 142
house numbers, 142
housing, 163
Hugging Face, 105
human annotation, 74, 77–8, 258, 274–6, 287–8, 290–1
human-in-the-loop, 72, 308
Humanities and AI Virtual Institute (HAVI), 307, 312

idiographic approaches, 14, 310, 314
images, maps downloaded as, 105
image segmentation model, 44–6, 145
imperialism, 20
incomplete, data is always, 36. *See also* missing data
index maps, 61–3, 65, 68, 69–72
individual choices, 135, 148, 165
individual-level data, 139, 140, 179, 184–97
industrial accidents, 161
industrialisation, 3–4, 157, 181–2
Industrial Revolution, 2, 6–26, 131–3
inference results, 105–6, 115, 116, 118, 146
information society, 3–4
information theory, 56
Integrated Census Microdata (I-CeM), 18, 23, 24, 139–40, 171, 184, 243, 244
interactive Notebook, 259, 282, 290
inter-annotator agreement, 289, 290, 291, 292
interdisciplinarity
  computational map studies, 91, 106–7, 108
  convergence experiments, 132
  disciplinary interstices, 22
  language of mechanisation, 257, 289
  versus multidisciplinary, 21
  processes of, 23
  project based on, 4–5, 305–7, 314
International Image Interoperability Framework (IIIF), 87n74, 103, 105, 109
interpretability, 308
'interrupted' datasets, 74. *See also* missing data
inventors, 8
Ireland, 20

Jackson, Alan, 150, 151, 152
Jensen-Shannon Divergence, 56
JISC, 16, 17, 37, 55–6, 57

Johnson, James, 144
Johnson, Jessica Marie, 306
journalism, 13
Joyce, James, 135, 169
JSON, 74, 75, 105
Jupyter Notebooks, 74, 99, 101, 291–8, 315

Kellett, John, 134, 135, 136
keyword-based annotation, 78
keyword frequencies, 74
keyword search, 34, 38–9, 59
King's Data Lab (KDL, King's College London), 276, 277, 278, 279, 280, 283, 284, 285, 286, 291, 295, 296, 297
Kitchin, Rob, 36
Klein, Lauren F., 39, 57, 60, 99, 306, 308, 309
k-means, 265, 268

labelling functions, 78
labelling maps, 94, 97, 98, 100, 104–5, 114, 143
land access, 137
landscape comparison, 55–8, 65, 74–5
landscape-map relationship, 96, 112
land use data, 93
large language models (LLMs), 89n88, 310, 311
Late Modern English, 255
latitude/longitude coordinates, 66
layers of (map) data, 93, 97
Leeds, 146
legal framework, 138
leisure rail travel, 135–6, 150
lemmatisation, 260
liberalism, 9, 52, 53
Library of Congress, 107
LiDAR, 117
life expectancy, 10
light industry, 158
Lilley, Kirsty, 116, 117
line geometries, 142, 145
Linked Open Data, 47, 48, 97–8
'Linker,' 48
literacy, 256
literary studies, 112, 256–7
Liverpool Street, Salford, 165
loading maps and metadata, 103
local historians, 139
local interest newspapers, 57
local nature of railways, 132
London. *See also* Camberwell; Upper Holloway
  cycle trade, 182
  rail infrastructure, 133, 134, 137
Long, Hoyt, 310

machine learning
  annotation, 77
  fine-tuning models, 44
  MapReader, 100, 118
  maps, 65, 95, 98
  Microsoft Books, 74
  modelling alternate worldviews, 61
  novel research instruments, 38
  post-processing, 47
  US Census Tree, 187
machine-readability, 4, 24, 43, 145
Machines Reading Maps (MRM), 94, 107–8
Macromap, 68–72, 101
Maker-Dealer sector, 198, 220, 225, 226
Making of the English Working Class (Thompson, 1963), 3, 10
Manchester, 133
Manifold platform, 6, 25
manual annotation, 47, 105, 107, 117, 146, 286, 289–98
manual corrections to data, 47
manual record linking, 187
manual transcription, 46
Manufacturing sector, 198, 202, 203, 220–2, 223, 226, 231–2, 235, 245
map collections, 65–6, 95, 144
mapKurator, 107
Mapping Inequality, 117
mapping the landscape, 49–55
MapReader, 22, 24, 65, 72–3, 110–16, 139, 144–7, 169, 317
'Map Records search,' 66
maps
  bias, 144
  critical search, 67–9
  as data, 11, 96, 139, 313
  Environmental Scan, 61–73
  formats, 102–3
  patches, 92, 94, 96–9, 102–18, 145–7
  railway data from, 143–7

maps (*cont.*)
revisions and reprints, 70
scales, 144
transcription of text, 73
map scales, 67, 92
map sheets as collections, 63–5, 92–3
maptext, 92, 94, 107–8, 116, 118
marginalised voices, 60, 80, 309
market forces, 137
Martin, Meredith, 306
Marx, Karl, 7
mass market newspapers, 55, 56
'match' semantic change, 272, 273
McDonough, Katherine, 24, 107, 109, 116, 118
McGillivray, Barbara, 274, 283, 284, 285, 286, 292
meaning expansion, 267
mean railspace and distance to station, 157–8
mechanisation, 255–303
mega-cities, 169
metadata. *See also* enhanced metadata
book collections, 74–9
creating from reference works, 43–7
critical search, 67–9
crowdsourcing, 77–8
Environmental Scan, 33–4, 36, 38, 39, 40
Google Books, 13
historical reference metadata, 40
machine annotation, 77
'Macromap' visualisation, 68
map collections, 65–6
map patches, 105
maps, 101, 145
newspapers, 260, 272–4, 290, 309
rail infrastructure, 72
railway passenger stations, 147
series maps, 61
metalworking, 180–4, 197–8, 202–3, 213, 224, 227–8, 231, 235, 240, 241, 248
microcensus data, 113, 139, 140
Microsoft Books, 37, 74, 77, 78
middle classes
amenity and disamenity, 152, 154
choices about where to live, 135, 148
rail infrastructure, 138
Salford, 167
surrogate markers for wealth/income, 148–9
Middlesbrough, 156, 157, 158, 163–4
Midlands, 180, 181
Midland Works, 158
migration, 184, 213, 240
Mill, John Stuart, 2
Mimno, David, 306
'minimum research outcomes', 22–3, 315
Miro boards, 296
mirror worlds, 96, 111
missing data, 36, 37, 39, 40, 56, 74–5, 101, 309
missing word prediction tasks, 311
mistranscription, 39
Mitchell's Press Directories, 17, 41, 42–3, 48, 49, 54, 55, 290, 308
modelling alternate worldviews, 59–61
modern roads dataset, 140–1
monotone-x curves, 278
Moor, Edward, 276
Morris, Victoria, 77
motor vehicle industry, 183, 185, 197, 230–1, 232, 245
Muehrcke, Philip, 110, 111
multidimensionality, 5, 15, 39, 94, 261, 309
multidisciplinarity, 21
multimodal cartography, 94, 108
multimodality of knowledge, 312–13
multiscalar research, 92, 95
music streaming, 33

name changes, 193–4
named entity recognition (NER), 83n26
National Archives, 18
National Grid References, 66
National Library of Scotland (NLS), 19, 37, 63, 65–9, 72, 92, 101, 107
national mapping projects, 63
Natural Language Processing (NLP), 83n26, 257, 258
nearest neighbour sets, 261, 263, 265, 267, 269, 270, 271, 272
newspaper archives, 13, 16. *See also* British Newspaper Archive (BNA); digitised newspapers
Newspaper Press Directories, 41–4, 47–8, 49–54, 58–9
newspaper publishing, 3, 4, 11, 15, 40

Newspaper Word Count Corpus, 58–9
new workers, recruitment of, 198, 202, 203, 220, 240
Ngram viewer, 12
noise pollution, 148, 156–7
noisy/messy data, 257, 264, 289
nominal linkage method, 192–3, 194, 220, 240, 241–2
nomothetic approaches, 14, 310, 314
non-relational data structures, 98
Norcliffe, Glen, 183, 184
Notebooks, 259, 276, 277, 282, 290, 291–8
nuisance, 137, 148

object detection models, 145
Observable notebooks, 69, 259, 277, 282, 290, 291–8
occupational data
  ancestors' occupations, 220–9
  census, 18, 143
  cycle trade project, 184–97
  household head, 220–9
  occupational destinations, 229–43
  occupational origins, 197–229
  railspace, 148–9, 152–3, 159, 171
Ogburn, William, 3, 4
Old Kent Road, London, 161
oligoptic data, 36
O'Malley, Tom, 41
onomasiological analysis, 283
open-access datasets, 107, 142, 289, 317
Open Maps, 95
open source code, 45, 93, 99–100, 109, 116, 169, 257, 259, 316–18
open-source spatial data, 24
optical character recognition (OCR), 13, 34, 44–6, 59, 107, 261–4, 289
Ordnance Survey
  as data source, 19, 20, 37, 144
  digitisation, 67
  GB1900, 142–3
  and industrialisation, 3, 24
  metadata, 65–6, 101
  OS Open Roads, 24, 140–1, 142–3
  railspace, 138–9
  series maps, 61–2, 64–5
  surveys, 92
  topographic maps, 144
  Town Plans, 144
outliers, 168–9
overcrowding, 163, 168
oversampling, 56, 59, 75, 76, 80
overwhelm (information), 34, 91, 92
Oxford English Dictionary (OED), 257, 265, 270, 275, 276, 277, 281, 282, 287

paper and ink quality, 59
parish-level census data, 140, 142, 143, 158
partisan labels, 49–54, 56–7, 58
patches, 92, 94, 96–9, 102–18, 145–7
paywalls, data hidden behind, 4, 5, 316
Pedrazzini, Nilo, 274
Pelagios, 98
Pemberton Gardens, Upper Holloway, 159, 160
Pepys, Samuel, 281
personal liberation, 11
perspective of sources, 39, 57, 60, 144
pixel-level labelling, 94, 96, 102, 105
place names, 48, 66, 108, 142. *See also* street name data
plant images, 107
plural words and semantic change analysis, 273–4
point geometries, 142, 143
political orientation of newspapers, 49–54, 56, 58, 291
pollution, 136, 137, 138, 148
Pooley, Colin, 134, 144
Posner, Miriam, 99
Post Codes, 66
post-processing, 47, 106, 262, 289
poverty, 11, 135, 136, 148, 256
precision, 98, 142, 143, 145, 278
prediction, 105, 311
pre-processing of images, 44
pre-processing of text, 260
press directories, 35, 40, 41, 42, 44, 47, 48
price point variables (newspapers), 55, 59, 73
printing, 13, 256
probability matching, 48
professional occupations, 143, 154, 155, 159. *See also* central business district workers; middle classes
project aims, 21–2
project labs, 22–3
provenance of digital data, 15, 39, 308–9

provincial newspapers, 15, 35, 41, 54, 55
proximity to rail infrastructure, 147–8, 151–2
public domain data, 139
Public Record Office, 18
Putnam, Lara, 34
Python tools, 60, 100, 113, 140, 185

QGIS, 93, 113
quantitative methods, 12
Queensland Road, Upper Holloway, 160

radical collaboration, 93, 118, 257, 314
radical print culture, 10, 11
railspace, 24, 99, 106–8, 113–16, 138–9, 145–7, 150–6
railway passenger stations, 139, 147
Railway Passenger Stations in Great Britain: A Chronology (Quick, 1996), 24 147
'railway' semantic change, 266–7, 269, 270
Ramsay, Stephen, 111–12
'Ranker,' 48
raster data, 93, 97, 145
reasonable match assessments, 194
recency bias, 13
reciprocity networks, 135
Recogito annotation tool, 107
'Recogniser,' 48
record linkage, 47, 48, 185, 189–97, 241–2
reframing analysis, 58–9
'Regent Road' parish, Salford, 157, 158, 165
regional industrial specialisations, 10, 181–2, 183
regional linguistic variation, 58–9, 258, 272–4, 287, 288, 289
regional under-representations, 56
remote sensing data, 117
representativeness, 35, 39, 55, 59–61, 74, 143, 306
reproducibility, 81, 169, 241, 242, 292, 316–18
residential differentiation, 134, 135
residential streets, 115
reuse, building for, 316–18
Rhodes, Joshua, 23–4
Roberts, Robert, 157, 158, 165, 167
Rostow, Walt, 7–8
Rowland, Edwin, 310
Rowntree, Seebohm, 148–9, 150
Royal Atlas of England (Bartholomew, 1899), 134
rural areas, 142, 143, 144, 149, 152, 169–70, 256

Salford, 156, 157, 158, 165–7, 168
Sample Inspector, 74–5
sample sizes, 195, 243
sampling strategies, 60
Sanborn fire insurance maps, 87n67, 117
sanitation, 136
Sankey plots
  construction of, 197–8
  examples of, 199–201, 204–6, 207–12, 214–19, 233–4, 236–9
scale as analytical question, 14, 132, 241
scanning, as part of digitisation process, 43–4, 65, 68, 93, 96, 101, 102
Schmidt Sciences, 307
Schurer, Kevin, 195–6, 199–201, 204–6, 207–12, 214–19, 221, 224, 227, 230, 233–4, 236–9
SciVision, 22, 107, 117
Scotland, 177n59, 185
searching digital collections, 33–5, 38–9, 93
selection criteria for digitisation, 74
selfhood, and social class, 10
semantic annotation, 47, 257, 258, 287–8, 290–1
semantic change detection research, 260, 310–11
semantic segmentation methods, 96
semi-manual processes, 48
sense disambiguation tasks, 258, 274–6, 287–8, 290–1
sentiment analysis, 13
series maps, 61, 92, 142, 144
servant-keeping, 148–9, 151–2, 159, 160–6, 168
sewing machines, 182, 185, 197, 220
Shalizi, Cosma, 305–6
shapefiles, 67
silk weavers, 182, 222, 223, 226
Simon, Rainer, 116
Singer, George, 180
situated knowledge, 310–12

situated methods and models, 310–12
slavery, 9, 12
slum clearances, 136
slum development, 156, 157, 161, 165, 167
small-multiples-inspired visualisations, 68, 101
Snow, C.P., 4
So, Richard Jean, 310
social class. *See also* central business district workers; middle classes; working classes
  linguistic variation, 259
  marginalised voices, 60, 80, 309
  and railway infrastructure, 24
  and selfhood, 10
  servant-keeping, 148–9, 151–2, 159, 160–6, 168
social group-level census data, 179–80
social history, 2–3, 10, 132
social inequalities, 80
social obligation networks, 135
social reform agendas, 9
socio-economic representativeness, 143
sociology, 134
software communities, 106–9
Software Sustainability Institute, 109
source criticism at scale, 35, 36, 37, 38, 40, 79–81, 101, 308
source criticism, 61
spatial and social groupings, 134
spatial datasets, 138–47
spatial humanities, 93
spatial search, 66
spelling variations, 192, 262, 264
'stamp' semantic change, 272–3
Standard English, 256
standardisation processes, 140
standards of living, 10
Stanford Libraries, 107
Stanford Literary Lab, 40
Starley, James, 180
Starley Brothers, 182
statistical analyses, 10, 14, 54, 278, 279
statistical bias, 56
steam engines, 2, 8, 136, 138
StopsGB, 24, 139, 147
Strabo, 107
street-level data, 138–47, 156–67, 168, 313
street-level versus house-level data, 142
street name data, 140, 142
streetscapes, 150–6
structural topic models, 53
subcorpora, 272–4
sub-regional industrial specialisation, 181–2
subsamples, weighting, 60, 80, 309
subsamplr, 60
sub-teams, 23
suburban growth, 133, 150, 156, 165
suburban rail, 169
Sultan Street district, 156, 161, 163
surrogate markers for wealth/income, 148–9, 158

tabular data, 44, 47, 48
team annotations, 114
technological convergence, 197, 241
technological determinism, 9
technological vocabulary, 47, 58–9, 255–303, 312
temporal dimensions, 69–72, 169, 179, 312
temporal evolutions, studying, 49, 51, 102
temporal semantic trajectories, 58–9
temporal unity, 65
Tesseract 5, 45–6
textile industry, 9
Text Machine: Computing Literary Innovation, 312
text recognition models, 45–6. *See also* optical character recognition (OCR)
text segmentation, 47
text spotting, 94, 107–8
Thomas, Edward, 133
Thompson, E.P., 3, 10
tileservers, 67, 102, 105
Tilton, Lauren, 306
time-series analysis techniques, 265, 287
'time slices,' 260, 264, 312
Tolfo, Giorgia, 24
topic prevalence studies, 53–4, 75
toponym linking, 48
town planning, 136
Toynbee, Arnold, 7, 8, 9, 133
training models, 38–9, 47, 78, 79, 94, 114, 258, 260, 311
'train' semantic change, 266–7, 269, 287
trams, 169, 280
transcription, 73

transferability, 316–18
transfer learning, 83n26
transparency, 74, 80, 100, 242
transport datasets, 140
T-Res, 48
'trolley' semantic change, 276–80, 288, 292
Turing Way, The, 100
typefaces, 44

Ulysses (Joyce, 1922), 135, 150, 169
under-sampling in data collections, 56, 75
UNet, 44
unique IDs, 105
unique personal fingerprints in census data, 188
University of London Press, 6, 25, 317
University of Wisconsin, 110
Upper Holloway, 156, 157, 158, 159, 160, 168
urban history, 132, 134, 136, 144
Urban History Group, 134
urbanisation, 131, 133, 136, 169
urban maps, 144
urban poor, 136
US, 183, 185, 187. *See also* American English
useful friction, 100
user interfaces, 97, 100, 296
utility, 137

Vane, Olivia, 69–71, 101
van Strien, Daniel, 24
vector data, 93, 98, 99, 112, 113, 145
Venice Time Machine, 117
versioning, 100, 315
visual indexes, 68

wages, 9–10
Wales, 142
Walsh, Melanie, 306
waste disposal, 163
watch and clock makers, 220–1, 223, 226, 227
wealth/income data, 148–9
'we' and 'they' language, 58–9
Webb, Beatrice, 133
West Midlands, 180–97, 198
White, Jerry, 161, 163
'wide scan', 39, 40, 54, 55, 61–79, 309
Wikidata IDs, 48, 73
Wilson, Daniel C.S., 24, 116
Windelband, Wilhelm, 14, 310
Wind in the Willows (Grahame, 1908), 284
'winnowing down,' 69
women's name changes, 193–4
Wood, Rosie, 116
Word2Vec, 260
word count corpora, 58–9
word embeddings, 258, 260–4
word usage frequency, 12
word vector representations, 260–4
worker retention, 202, 203, 235
working classes
  dangers of rail proximity, 157
  employment, 155
  newspaper publishing, 16
  radical print culture, 10, 11
  rail infrastructure, 136, 138
  standards of living, 10
workshops, 109
World Historical Gazetteer, 73
worldviews, modelling alternate, 60–1, 150

XML, 74
Xylonite collar works, 160–1

young workers, recruitment of, 198, 202, 203, 220, 229, 240

zoomability, 67, 68–9, 103
Zooniverse, 77, 78, 275, 276, 290–1, 292, 293, 295