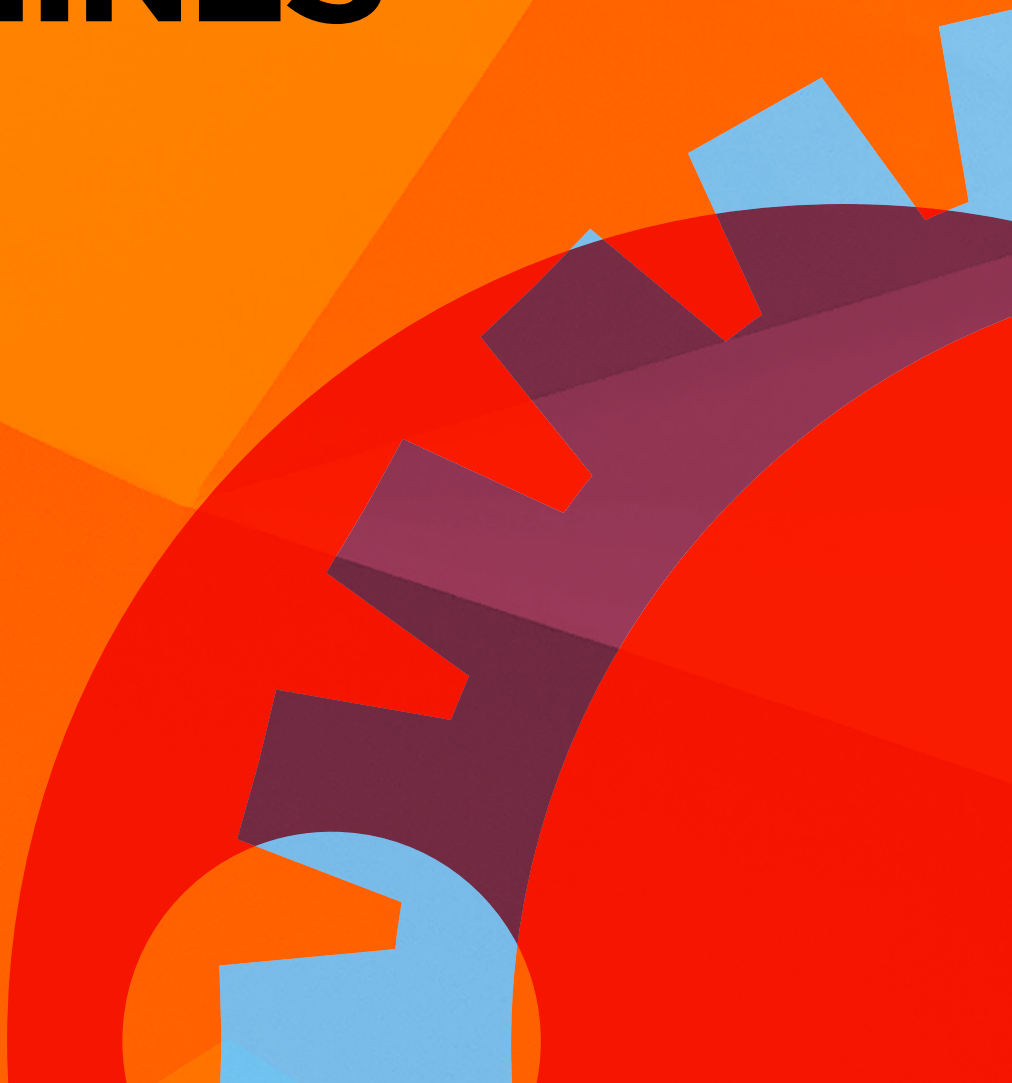
LIVING WITH MACHINES



UKRI Arts and Humanities Research Council

Living with Machines is both a research project and a bold proposal for a new research paradigm. In this ground-breaking partnership between The Alan Turing Institute, the British Library, and Universities of Cambridge, East Anglia, Exeter, Queen Mary University of London, and King's College London, historians, data scientists, geographers, computational linguists, library professionals, and curators were brought together to examine the human impact of industrial revolution. Living with Machines was funded by UK Research and Innovation (UKRI), via the Strategic Priorities Fund, and was administered by the Arts and Humanities Research Council (AHRC).

Contents

Principal Investigator's Introduction

It is with great pride that I write this end of project report, as well as some sadness. When the other investigators and I set out the vision for this project in 2017, we had some big dreams. Living with Machines was imagined at once as a data-driven history project, and a historically-informed data science project. Our object of interest was an earlier moment of huge technological upheaval: the coming of the machine age in the nineteenth century. One consequence of industrialisation in Great Britain was an explosion in the creation and collection of information. Our aim was to show how, thanks to decades of digitisation, we were now in a position to leverage this information as data through computational means. We proposed to develop innovative computational approaches to facilitate a new kind of history that would allow us to tell the stories of the impact of mechanisation on the lives of ordinary people.

But that historical aspiration was just one dimension of what we hoped to achieve. More fundamentally we wanted to develop computational models, tools, code, and infrastructure that would be transformative to the future study of cultural heritage collections. For a project to achieve such broad ambitions, we needed a diverse set of skills and expertise. Over its lifetime, the project has brought together historians, data scientists and research software engineers, curators and library professionals, computational linguists, digital humanists, visualisation experts, literary historians and an urban geographer. As such, Living with Machines was a bold experiment in radical collaboration across disciplinary areas and professional spheres. The experience of building a team and united project vision from these disciplinary parts has been an important experience in its own right, and reflection on this process has led to the publication of our first project book, *Collaborative Historical Research in the Age of Big Data: Lessons from an interdisciplinary project* (Cambridge University Press, 2023). It is important that others can benefit from our experience, and our challenges. Building a team such as this means that we not only draw expertise from a breadth of different communities, but develop a very different set of outcomes from those normally generated by history or data science projects. We are just as proud of the open source code and tools, and our databases and language models, as we are of our historical interventions. As our statistics and case studies will show in the following pages, we have been immensely productive in this space, creating 47 public Github code repositories to date, as well as several more polished tools that are being taken up by the community, such as MapReader, DeezyMatch, and T-Res. Through these tools we have created textual data from newspapers, visual data from maps, and tabular data from census returns, as well as digitising 488,000 additional newspaper, and over 15,000 new map sheets, in addition to other material such as the Road Acts and Micthell's Newspaper Press Directories.

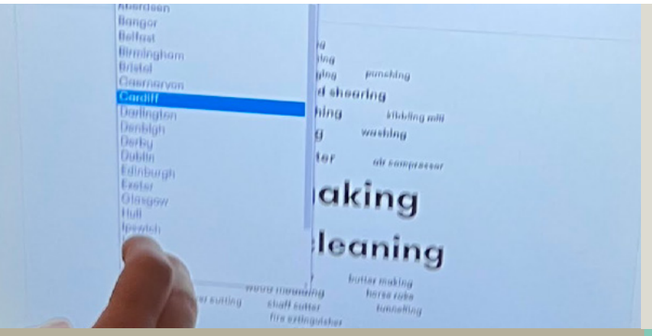
Thanks to working with the British Library, we have also been able to engage with the public in sustained ways throughout the project, not just at its end point. The project exhibition at Leeds City Museum – *Living with Machines: Human stories from the industrial age* – attracted over 42,000 visitors. The exhibition incorporated key themes and early outcomes from the project. Exhibition themes were also informed by discussions with the public during crowdsourcing projects about the impact of machines, and work on OS Maps was represented by an animation showing changes over time around Leeds. Crowdsourced material was also incorporated into two interactive visualisations. Beyond the exhibition, crowdsourcing has formed a central plank of several pieces of research on the project, including analysis of the language of mechanisation, and the study of industrial accidents - both of which will be featured in the project's second book, *Living with Machines: Computational Histories of the Age of Industry* (in progress). We are very pleased that our research has also reached the public through the development of our work on OS maps into a story run by *The Economist* in April 2023.

One of the aims of UKRI's Strategic Priorities Fund, which supported this project, is to 'drive an increase in high quality multidisciplinary and interdisciplinary research and innovation'. This is not simply about delivering some multidisciplinary and interdisciplinary research outputs, but driving a change in research culture beyond the project - for us, in the space of digital history and the use of cultural heritage collections. Where many endeavours fall short is with the attitude 'if you build it, they will come'. New methods, tools and datasets are of no use if nobody knows about them. In our final phase of the project we have been focusing on ensuring the legacy of our collaboration. As well as making our data and code open for reuse, we are seeking to develop communities of users around our most important tools and methods through blog posts, workshops, in-person and published tutorials. We have also encouraged the uptake of our methods and data through the award of six 'digital residencies'. These residencies are small fellowships or project awards designed to enable work around one of our datasets or tools. The fruits of their

labour will be reported in due course on our project blog (https://livingwithmachines.ac.uk/latest/). Finally, we are maximising our impact by creating a 10-part online docuseries. These short videos help us communicate with a broader audience about what we have been up to, why, and how our work might be useful to them.

In these final weeks of the project we are able to take stock and congratulate ourselves on the enormity of what we have achieved. As well as developing new computational methods we have all learned a lot about collaboration in the process, and we will take this into our future work. Perhaps the most exciting thing about this venture is that with all that we have built, we have only just begun to scratch the surface of what is possible. For this reason we are exploring opportunities to develop spin-off bids and projects. One has already been successfully undertaken, Machines Reading Maps (UK PI, Katherine McDonough), and we are also excited that two of our research associates will continue at the Turing in Research Fellowships that will allow them to continue working on Living with Machines data and questions. Further bids are at various stages of development. More broadly, we believe we have shown that library, arts and humanities scholars not only deserve a place at the table with those designing the future of AI and data science, but also have skills and knowledge that are vitally needed.

We are just as proud of the open source code and tools, and our databases and language models, as we are of our historical interventions.



42 MEMBERS

of the collaborative team over the lifetime of the project

488,000

newspaper pages newly digitised, and over 15,000 additional maps sheets

OVER 42,000

people attended the project exhibition

47 GITHUB

public repositories

OVER 700

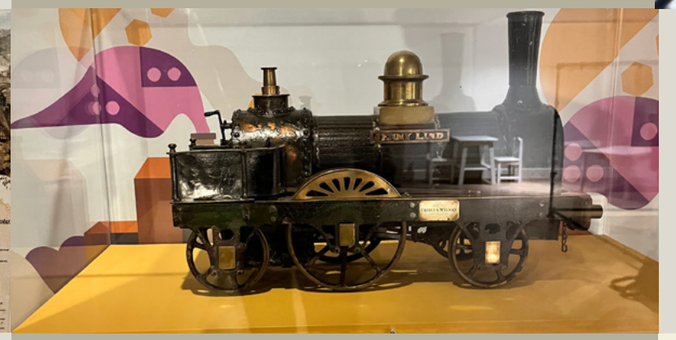
people reached and approximately 400 engaged via our Distributed Conference

OVER 190

workshops, talks and papers delivered

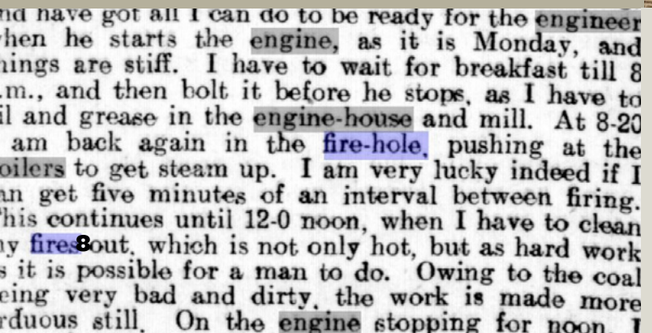
OVER 5,500

online volunteers from high school students to retired microbiologists



35 ARTICLES

books chapters and conference proceedings so far

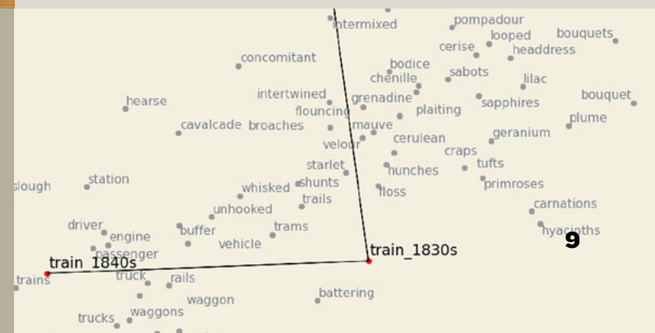


250,000

contributions to crowdsourcing tasks via Zooniverse

16 TEAM MEMBERS

transitioned into permanent or follow-on posts



Research

The interdisciplinary nature of Living with Machines, and the varied skills of the team that we have assembled, means that the project has been able to make contributions of various kinds. It has developed tools to make the digitised collections of historical documents research-ready by wrangling them into the accessible formats. By generating new datasets, as well as curating new derived or sample datasets from licensed material, we are enabling others to reproduce and build on our results. By releasing new contextual datasets we can now understand the contours of what has been digitised.

We have developed new tools and methods for, amongst many other things:

- Computationally 'reading' maps and identifying features such as rail and buildings (or anything else with a distinctive visual form).
- Extracting toponyms from text, disambiguating 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.
- Analysing the ways that that language evolved to describe new technological realities, specifically developing new frameworks for annotating word meaning in historical texts, and new algorithms for identifying changes in word meaning over time and across space.
- Studying the reporting of accidents involving machinery by enhancing the crowdsourced identification of newspaper articles discussing accidents involving machinery with customised machine learning classifiers.

The scale of our interventions are therefore expressed not only by the number of publications, but also the number of public Github repositories we have produced supporting our publications and tool releases, the number of datasets and pages of newly digitised material we have released, and the number of tutorials we have delivered or published.

This section of the report features just four case studies from the project. It was very hard to make these choices as there is so much rich work that has emerged. We have sought to provide a taster of the range of work by featuring case studies on (1) a tool, (2) a dataset and new method, (3) a method that generated a new historical account, and (4) an output sharing our reflections on the process of doing such research. To gain a broader perspective on how the work has come together to facilitate new historical insights and perspectives, we ask you to look out for our book in progress: *Living with Machines: Computational Histories of the Age of Industry* (forthcoming with University of London Press).

For our full list of publications, software packages, datasets, and other outputs, see livingwithmachines.ac.uk/achievements

Can Machines read Maps?

Living with Machines was fortunate to be given permission by the National Library of Scotland to use their world-leading digitised Ordnance Survey collections. MapReader emerged from a series of experiments on these maps using Machine Learning (ML) and Computer Vision (CV). It is a free, open-source software library written in Python for analysing large digitised map collections. This library transforms the way researchers can use maps by turning extensive, homogeneous map sets into searchable primary sources. MapReader thus aims to normalise using digitised maps at scale in the humanities and beyond.

MapReader allows users with little or no CV or ML expertise to i) retrieve maps via webservers; ii) pre-process and divide them into 'patches'; iii) annotate patches; iv) train, fine-tune, and evaluate deep neural network models; and v) create structured data about the visual elements of map content. As a result of our early experiments, we open-sourced c.62,000 expertly annotated patches with the hope that this dataset will foster further collaboration between the fields of CV, machine learning, history, and human geography as well as with libraries and archives.

In the first application of MapReader, we focused on British rail infrastructure and buildings as depicted in a collection of more than 16,000 nineteenth-century British Ordnance Survey maps (containing more than 30.5 million 'patches'), demonstrating how patches on their own and in combination with other patches or external datasets (such as another new, open LwM dataset of British passenger stations) can offer new insights into deciphering the spatial patterns of industrial development in modern Britain. Using an image classification task at the 'patch' level transforms a common, indeed unsophisticated, CV method into a radically new way for researchers to interact with maps. Patches are small, user-defined regions of maps implemented by overlaying a grid of equally-sized boxes across the scanned image. Interrogating the national landscape in terms of attributes like the footprint of the rail network or building density sets the stage for CV-driven research into a wide range of other important features of the built and natural environment captured by historical maps. The implications for informing ecological interventions to mitigate the climate and biodiversity emergencies and urban and rural trends in industrialisation, deindustrialisation, and regeneration are particularly exciting.

MapReader and its 'patchwork method' can be applied to a wide variety of problems in Digital Humanities and beyond. Indeed, we recently started a new collaboration with the 'scivision' team at The Alan Turing Institute to apply this method to plant phenotype images. We have also been approached by major private map collector David Rumsey (whose maps are now housed in the David Rumsey Map Center at Stanford University Libraries, forming one of the largest digitised

collections of historical maps in the world) to scope a collaboration building on the MapReader work and the LwM spin-off research project Machines Reading Maps (which is developing automatic methods for capturing and geo-locating text on maps).

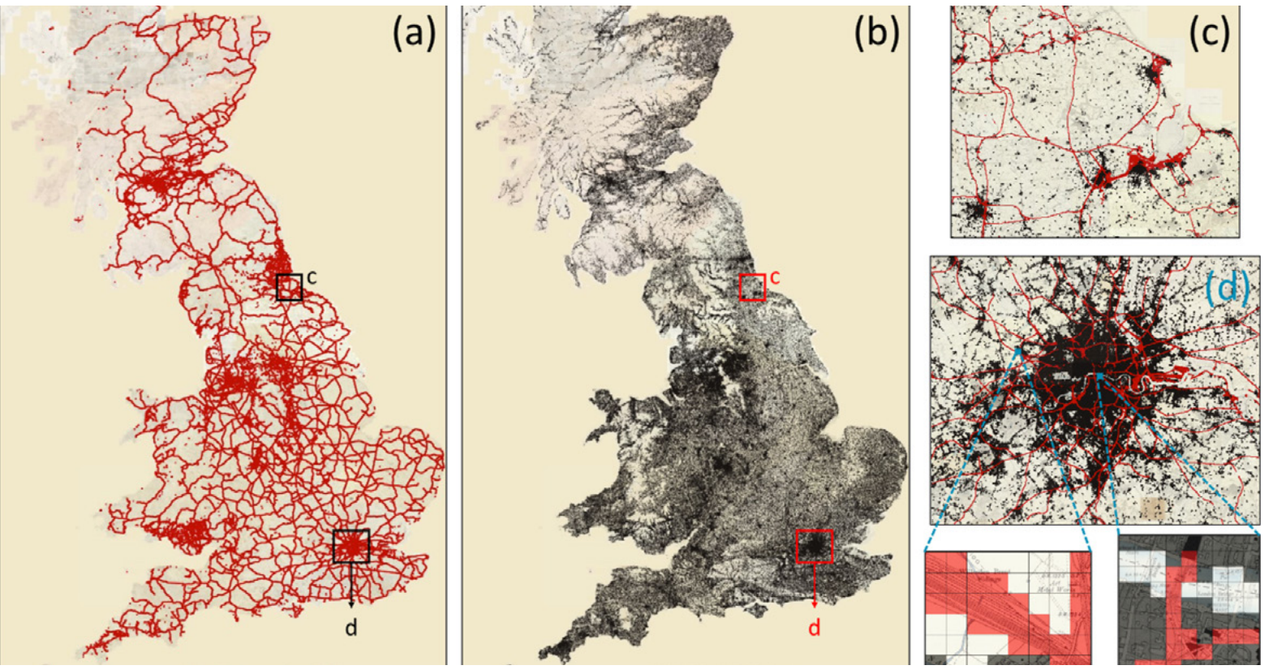
"MapReader transforms the way researchers can use maps by turning extensive, homogenous map sets into searchable primary sources."

Repository: https://github.com/Living-with-machines/MapReader

Documentation and Tutorials: https://mapreader.readthedocs.io/en/latest/

Introduction to the 'patchwork method': 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*, Volume 26, Issue 2, April 2021, Pages 284–299, https://doi.org/10.1093/jvcult/vcab009

Technical Paper: Kasra Hosseini, Daniel C. S. Wilson, Kaspar Beelen, and Katherine McDonough. 2022. 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, New York, NY, USA, 8–19. https://doi.org/10.1145/3557919.3565812 https://doi.org/10.1145/3557919.3565812



Tackling Bias: Introducing the digital 'Environmental Scan'

Digital collections are often opaque pools of information. Finding out what materials they contain is a non-trivial exercise; determining representativeness—and how this may affect analysis and interpretation—is even more difficult. However, contemporaneous reference sources can hold the key, contextualising digital collections and enabling us to compare digital collections with the wider information landscape from which it was generated.

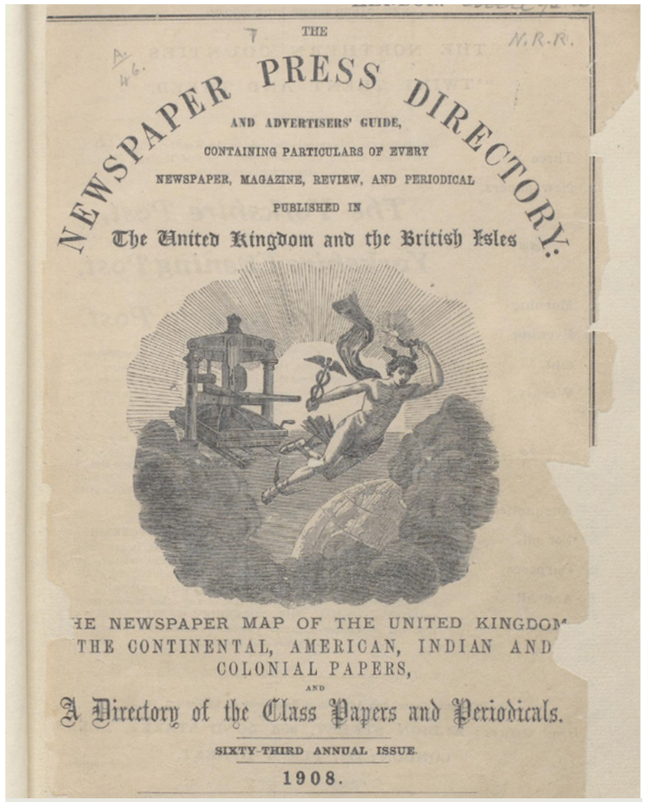
On Living with Machines the particular question we wanted to answer was how representative are digitised newspaper collections (such as JISC-digitised newspapers and the much larger British Newspaper Archive) of the wider Victorian press. Our implementation of the digital Environmental Scan on these collections began with the project's digitisation of a contextual publication, Mitchell's 'Newspaper Press Directories' (NPDs), the oldest in its genre, which was published almost yearly from 1846 onwards. Mitchell's Directories provide a rich description of each British newspaper, detailing its price, ownership, places of publication and self-ascribed political leaning, character and principal audience. We use this resource as a 'Victorian perspective' on our digital collection to address questions of representativeness. This dataset is now available for others to use via the British Library repository [link below].

We computed the extent to which our digital collections (the sample) resembled or diverged from the wider newspaper landscape (the population) as captured by Mitchell's Directories. Our main finding from our first study was that JISC radically over-sampled higher priced and party political newspapers and under-represented cheaper and less partisan ones. In our follow-up paper (in progress) we develop this method in two ways. Firstly, we use the digital Environmental Scan to explore how the representativeness of Britain's main digitised newspaper collection has changed over time since JISC and the British Library began building a digital corpus in 2004. Secondly, we investigate how the metadata derived from the press directories corresponds with newspapers' actual printed content. What distinguished the vocabulary of partisan newspapers, both 'liberal' and 'conservative', from that used by other newspapers? How did the language used by self-described 'independent' newspapers differ from that of newspapers calling themselves 'neutral', and so on?

We show that Conservative newspapers were distinguished by their extensive coverage of two things: the Church of England and agriculture. The most distinctive feature of Liberal newspapers was their coverage of religious nonconformity, the political system and temperance. Independent newspapers gave more coverage to both politics and religion than neutral ones, and were distinctive in their use of the vocabulary of balance and independence. Perhaps most striking of all, we show that the problem of poor OCR quality (the mistranscription

of printed words during the automatic text transcription process) is not random. The lists of distinctive words generated for more expensive and for Conservative newspapers are almost all real words, whereas the lists generated for cheaper and for Liberal and neutral newspapers are dominated by OCR errors (i.e. non-words). This is likely to be a consequence of cheaper newspapers being printed on poorer quality paper. It is an additional factor to bear in mind both when we manually search a newspaper database, and when we analyse newspapers at scale.

The Environmental Scan method demonstrates that even very large data sets contain hidden biases that shape how we see the past. It provides us with a means to contextualise our findings when we search or analyse the digital press, and it enables us to address these biases systematically by interrogating how the content of historic newspapers differs according to their political affiliation, price, place of publication and much else besides. The Environmental Scan also gives us the means to create our own bespoke sub-samples from existing digitised collections, and to inform future digitisation strategies. Perhaps most important of all, it is a method that can be reproduced for any source type in any country; all it requires is the survival of contemporaneous reference works from which we can generate new metadata - that is, new contextual information that enhances our critical understanding of digital sources.



Repository: 'The Newspaper Press Directory (1846-1920) - enriched and structured version', https://doi.org/10.23636/pbq5-9k28

Introduction to method: Kaspar Beelen, Jon Lawrence, Daniel C. S. Wilson, David Beavan, 'Bias and representativeness in digitized newspaper collections: Introducing the environmental scan', Digital Scholarship in the Humanities, 38: 1 (2023), 1–22, https://doi.org/10.1093/llc/fqac037

The Machine is Alive!

Machines have long straddled the fuzzy boundary in our imagination between lifeless artefacts and living beings. As such, they confound the usual linguistic markers of animacy, and are often portrayed as if alive: "the machine moved rapidly through the night!"! Such expressions – in which animacy, or life, is given to a machine – may be unconscious perceptions, or deliberate, through the use of figurative language. Writers sometimes choose to breathe life into inanimate objects, but either way; there can be consequences which are not only linguistic but also ethical and political as well.

The question of animacy has been studied by linguists, but existing computational approaches to detecting animacy neither aimed nor succeeded in atypical cases, such as machines. We developed an approach that built on recent innovation with neural language models. In particular, we took a BERT model and fine-tuned it on 19th century books from the British Library, producing a new model we call BLERT. Using an unsupervised approach and word embeddings, we were able to capture more fine-grained, contextual properties of words, providing a substantially more accurate characterization of both typical and atypical animacy, which can now be applied to large collections of digitised texts. Our results were published in *COLING*, a top venue in computational linguistics.

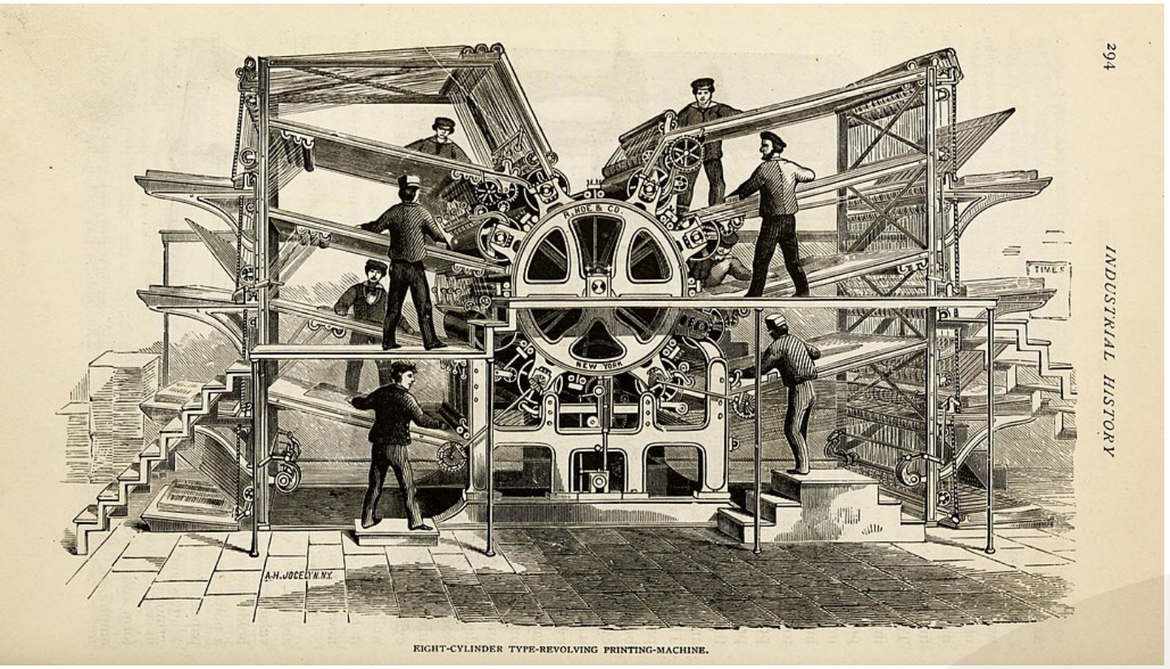
Extending this approach using a masked language model – a sort of a game in which we ask a model to predict likely words to complete a sentence – we developed this method further, expanding its scope: both to explore a wider set of texts, and also by pushing the boundaries of how language models can be used for humanistic research. Our method allowed us to detect unusual

"Humans and machines have always interacted in complex ways, which were accompanied by a changing language of descriptive figures and metaphors."

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy, 'Neural Language Models for Nineteenth-Century English', *Journal of Open Humanities Data*, 7 (2021), p.22, DOI: https://doi.org/10.5334/johd.48

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, Barbara McGillivray, 'Living Machines: A study of atypical animacy', *Proceedings of the 28th International Conference on Computational Linguistics (2020)*, DOI: 10.18653/v1/2020.coling-main.400

Daniel CS Wilson, Mariona Coll Ardanuy, Kaspar Beelen, Barbara McGillivray, Ruth Ahnert, 'The Living Machine: a computational approach to the language of technology', *Technology and Culture* (forthcoming July 2023)



or surprising turns of phrase, which we collected together into sub-corpora of relevant sentences, which we then close-read, providing detailed context and historical interpretation. Making use of the open-access 'Heritage Made Digital' newspaper collection digitised by the British Library, allowed us to explore these phenomena across genres of writing not always included in cultural and intellectual history, while conforming to best practice in reproducible research. This work will shortly be published in a long research article in *Technology and Culture*, the world's leading journal for historians of technology.

Radical collaboration

One of the most valuable outcomes of Living with Machines may not have been the research outcomes described above, but the *experience* of doing such research as part of a large interdisciplinary team. Interdisciplinary collaborations in the humanities happen, but it is not common. However, a number of forces have combined in recent years to make the convening of larger, and often interdisciplinary, teams a compelling and even necessary route for the study of our culture and history. While scientists in various subfields learn about how a lab is run from early on their research careers, and cultural heritage professionals have a range of complex collaborative projects to draw on, by contrast historians, linguists, literature scholars, and other humanists who find themselves involved in large projects or collaborative initiatives often do not have a blueprint to look to. This means that new projects and initiatives expend a lot of energy in their start-up period trying to establish collaborative values and project management strategies, often reinventing the wheel in the process.

As humanities research moves in this direction, it is increasingly important that we think deliberately and thoughtfully about how we design, structure, and undertake collaborative research. One way of doing this is by opening the doors on the internal workings of different projects. Our project sought to do this by writing a short open access book with Cambridge University Press, which provides a

We need to think deliberately and thoughtfully about how we design, structure, and undertake interdisciplinary collaborative research.

vision of how multi- and interdisciplinary collaborative research teams can work in ways that are greater than the sum of their disciplinary parts. To do so, it considered the pragmatic steps that facilitate intellectual exchange across those different disciplinary and professional contexts, which might be categorised broadly as the organisational issues of project management; legal and institutional issues of access to data; and the technical issues of hosting, wrangling, and analysing such data.

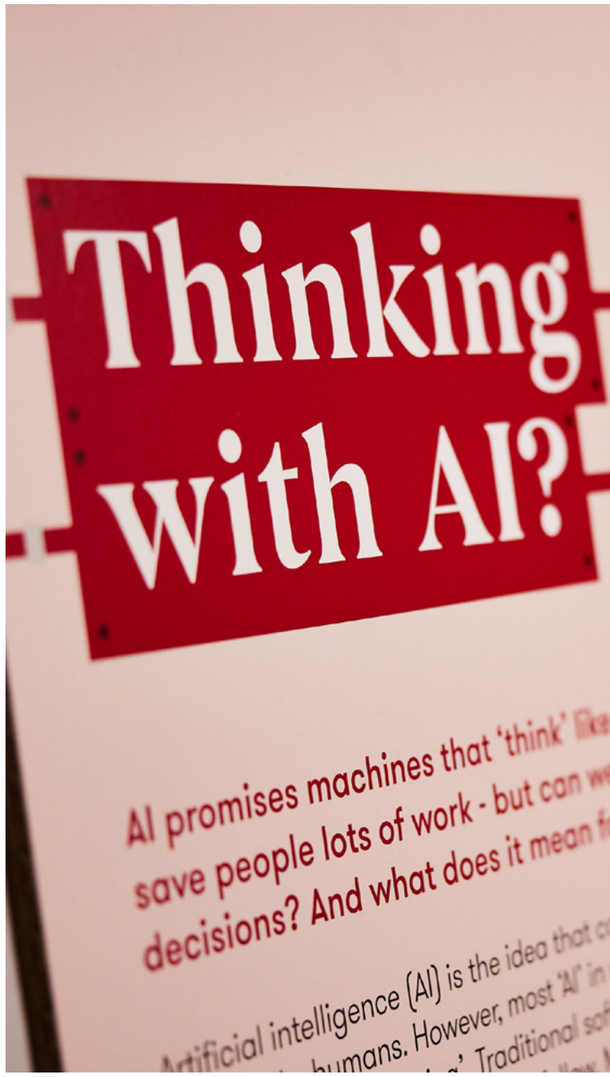
Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project* (Cambridge University Press, 2023), DOI: https://doi.org/10.1017/9781009175548

Events and Public Engagement

Project Exhibition

Living with Machines: Human stories from the industrial age was a free exhibition in the heart of Leeds at Leeds City Museum. A collaboration between the British Library and Leeds Museums and Galleries, the exhibition co-curators, Co-Investigator Mia Ridge and John McGoldrick, unearthed new stories behind British Library and Leeds Museums and Galleries collections to explore how machines and mechanisation changed how people lived and worked in Leeds, the local region, the Great Britain and its colonies. These stories were brought to life with artworks, ballads, machines, clothing, blueprints and banners and more on loan from the National Football Museum, National Railway Museum and many other institutions. From Dickens to Lowry to the victims of industrial accidents, women's football to striking workers, the exhibition painted a relatable picture of how rapid advances in technology changed life and work for everyone. Importantly, for the project, the exhibition made key themes and early findings and outcomes from the broader project accessible to the public.

The exhibition doubled our target, reaching over 42,000 visitors. A diverse programme of family and adult learning activities engaged a further 1,821 participants. Events included evening panels and a workshop that explored the implications of AI for the creative and cultural industries, and for businesses seeking to make ethical choices about AI.



Evaluation found that the exhibition successfully engaged a broad audience base, with 28% of visitors being residents from areas of socio-economic disadvantage. The exhibition's narratives and objects successfully engaged visitors with stories they could relate to to enhance learning and the relatability of the topic. For example, 80% could see parallels between the changes created by machines in the 1800s and those caused by technology now, and 67% said that the exhibition changed how they thought about the impact of machines on everyday life. 80% of visitors reported that it enabled them to empathise with the lives of people in the 1800s. Over half said that the exhibition changed how they thought about AI, an interesting contrast with the 80% who saw modern parallels with changes caused by machines. 51% also said that the exhibition inspired them to do something new. The exhibition also provoked ethical discussions on the role of machines and the themes of environmental impact, women's rights, health and safety and worker rights. Qualitative research found that prompting ethical discussions and reflection was a key impact of the exhibition. Importantly for us as a research project, 87% of visitors surveyed agreed or strongly agreed that the research behind the exhibition was important to be able to tell new stories and give new perspectives on history.

Distributed Conference

Most big projects culminate with a conference. Living with Machines chose to break with this convention in part due to the breadth of audiences with which we sought to engage, and in part inspired by the more expansive and creative forms of exchange that were opened up by the pandemic. Instead, during the final twelve months of the project we delivered a series of activities that allowed us to both to broadcast to and converse with colleagues with backgrounds in data science, libraries and information studies, humanities, policy and beyond.

Our first instalment in July 2022 included the online public event 'Humanity and Technology: In Conversation with Jo Guldi', then Professor of History at Southern Methodist University (US). Guldi delivered a breathtaking double bill of public lectures on the topics of 'Pseudo History and Digital History: The Dangerous Art of Text Mining', and 'The Long Land War: The Global Struggle for Occupancy Rights, 1881-1974'. The former included an important rallying cry for why data science needs to engage the humanities: the huge size and complexity of cultural heritage data has lessons for all.

Other events have included a series of workshops with invited participants to focus on specific outcomes from the project, and to scope additional opportunities and future challenges. These have included workshops on our tool MapReader, on the challenges of working with adverts in historical newspapers, on 'Machine Learning Approaches for Historical Trade Directories' (co-hosted with 'The Congruence Engine', of the AHRC's TaNC Discovery Projects), and on 'AI and Historical Newspapers' (co-hosted with the British Library). Such events have helped us to connect with the broader communities in the UK, in Europe, and in the US who are thinking about related data and research problems.

The project has also hosted three online roundtables on related themes of collaborative research on digital history projects - discussing candidly the benefits of such work, its challenges, and potential strategies. The first marked the launch of the project book, *Collaborative Historical Research in the Age of Big Data* (discussed in more detail above) taking the form of a conversation with the authors, and hosted by two of our wonderful advisory board members - who initially encouraged us to write the book - Professor Jane Winters and Professor James Smithies. The second, 'Digital History and Collaborative Research: a Practitioners' Roundtable' was co-hosted with the Royal Historical Society. Chaired by our PI Ruth Ahnert, it welcomed Daniel Edelstein (Stanford University), Maryanne Kowaleski (Fordham), Jon Lawrence (Exeter), and Katrina Navickas (Hertfordshire). The third panel 'AI beyond STEM: digital skills to unleash the power of data science and AI for all' chaired by David Beavan and featuring Kaspar Beelen (SAS,UCL), Nicole Colemane (Stanford), Mathilde Daussy-Renaudin (Oxford/UCL), Lydia France (Turing), and Katie Ireland (University of Georgia).

Finally we hosted two public launch events. The first was the MapReader Public Launch, a two-day event at which we heard from librarians, historians, and data scientists about how Living with Machines has experimented with National Library of Scotland (NLS) Ordnance Survey maps, and attendees had a chance to test MapReader with Ordnance Survey maps from the NLS and the BL. The event also benefited from two inspiring keynote lectures from Chris Fleet (NLS) and Nicole Coleman (Stanford University). Our capstone event was the double launch of this glossy report and the ten-part docuseries, directed by Léllé Demertzi. Both report and docuseries were designed to ensure the lessons from our project were disseminated to a wider audience. The ten episodes will be hosted on both The Alan Turing Institute's and the British Library's Youtube channels.

Team + Destinations

Ruth Ahnert, Principal Investigator, Queen Mary University of London.

David Beavan, Co-Investigator, Co-I (acting joint-PI December 2021-June 2022), The Alan Turing Institute.

Giovanni Colavizza, Co-I, The Alan Turing Institute. Departed June 2019 to take up a post as Assistant Professor of Digital Humanities at the University of Amsterdam.

Adam Farquhar, Co-I, British Library. Retired September 2019.

Emma Griffin, Co-I (acting joint-PI December 2021-June 2022) University of East Anglia.

James Hetherington, Co-I, The Alan Turing Institute. Departed January 2020 for a role as Director of Digital Research Infrastructure at UK Research and Innovation, before Director of UCL's Advanced Research Computing Centre.

Jon Lawrence, Co-I, University of Exeter.

Maja Maricevic, Co-I, British Library. Joined September 2019.

Barbara McGillivray, Co-I, University of Cambridge (to August 2021), and King's College London (from September 2021)

Mia Ridge, Co-I, British Library.

Sir Alan Wilson, Co-I, The Alan Turing Institute

Claire Austin, Rights Manager, British Library.

Kaspar Beelen, Digital Humanities Research Associate, The Alan Turing Institute. Departed April 2023 for role as Technical Lead, Digital Humanities, Institute of Advanced Studies, UCL

Mariona Coll-Ardanuy, Computational Linguistics Research Associate, The Alan Turing Institute.

Karen Cordier, acting Project Manager December 2020 to March 2021. Moved onto role as Fellowships Manager, The Alan Turing Institute.

Joel Dearden, Research Software Engineer, The Alan Turing Institute. Died July 2020.

Léllé Demertzi, Project Administrator, The Alan Turing Institute. Joined June 2022, and moved onto role as Programme Coordinator, Data Science for Science and Humanities, The Alan Turing Institute.

Rosa Filgueira, Data Architect, University of Edinburgh. Departed September 2019, back to her role as a Research Fellow at the EPCC (University of Edinburgh), and subsequently onto roles as an Assistant Professor at Heriot Watt, and Lecturer in Computer Science at University of St Andrews.

Sarah Gibson, Research Software Engineer, The Alan Turing Institute. Started May 2020 and left in May 2021 to take up a role as Open Source Infrastructure Engineer at 2i2c, an open source contributor and advocate.

Lucy Havens, Visiting Researcher, The Alan Turing Institute, June-September 2022.

Luke Hare, Research Data Scientist, The Alan Turing Institute. Joined July 2022.

Timothy Hobson, Research Software Engineer, The Alan Turing Institute.

Kasra Hosseini, Research Data Scientist, The Alan Turing Institute. Departed December 2021 for a role as Senior Applied Scientist at Zalando Technology.

Michael Jackson, Software Architect, University of Edinburgh. Departed in July 2019 back to his role as Principal Architect at the EPCC (University of Edinburgh).

Amy Krause, Data Architect, University of Edinburgh. Departed in March 2020 back to her role as Principal Architect at the EPCC (University of Edinburgh).

Christina Last, Research Data Scientist, The Alan Turing Institute. Started July 2021 and left in May 2022 for MIT where she is postgraduate student and US-UK Fulbright Scholar.

Sherman Lo, Research Software Engineer, Queen Mary University of London. Joined August 2022.

Katherine McDonough, History Senior Research Associate, The Alan Turing Institute (to December 2022), and Lecturer in Digital Humanities, Lancaster University (from January 2023).

Federico Nanni, Research Data Scientist, The Alan Turing Institute. Joined project November 2019 and left in May 2023.

Nilo Pedrazzini, Research Associate in Corpus-Based Digital Humanities, The Alan Turing Institute. Joined January 2022, and departed (partially) for a role as Research Software Engineer at the University of Oxford.

André Piza, Research Project Manager, The Alan Turing Institute. Moved onto role as Strategic Programme Manager, The Alan Turing Institute.

Griffith Rees, Research Data Scientist, The Alan Turing Institute. Joined June 2022.

Josh Rhodes, Research Associate, The Alan Turing Institute. Joined November 2020, and departed for a British Academy Postdoctoral Fellowship at the University of Durham in December 2022.

Affiliate Yann Ryan, Curator Digital Newspapers, British Library. Departed January 2019 for a role as postdoctoral research associate on the project Networking Archives.

Andrew Smith, Research Data Scientist, The Alan Turing Institute. Worked on the project in 2021 and again from January 2022.

Guy Solomon, Research Associate, The Alan Turing Institute. Joined April 2022, and departed December 2022 for a role as Postdoctoral Research Associate in Mathematics of Cities at the University of Glasgow.

Giorgia Tolfo, Data and Content Manager, British Library. Departed in February 2023 to join The National Archives as Collections Researcher.

Daniel van Strien, Digital Curator, British Library. Departed December 2022 for a role as Machine Learning Librarian at Hugging Face.

Olivia Vane, Digital Humanities Research Software Engineer, British Library. Departed November 2021 for a role as Interactive Data Journalist at The Economist.

Kalle Westerling, Research Software Engineer, British Library. Joined January 2022 and left the project June 2023 for a role as Research Application Manager at The Alan Turing Institute.

Daniel Wilson, History Research Associate, The Alan Turing Institute, and (since February 2022) also Research Associate on Towards National Collection Discovery project, The Congruence Engine, at the Science Museum.

Rosie Wood, Junior Research Data Scientist, The Alan Turing Institute. Joined in January 2023.

Digital Residents

In order to encourage a wider community of people to engage with our datasets and tools, we launched an open call for six digital residencies to support creative project proposals making use of them. The model follows the practice of libraries to fund academics to come and consult their collections. Since the pandemic we have seen more examples of such fellowships moving online, such as the National Library of Scotland's fellowships in digital scholarship or JSTOR's innovation fellowships. We had a large and strong pool of applicants and were delighted to award small grants to the following people:

Nicola Baldwin – to develop a dramatic response to our crowdsourced accidents data.

Nicolò Bonato – to develop an interactive web app to visualise the data contained in the Press Directories dataset about papers's political leanings, and pair it with historical results for general elections.

Jennifer Hayward, **Michelle Prain-Brice**, and **Leonor Riesco** – to experiment with alto2txt and other project pipelines to work on The Anglophone Chile News Archive.

Yann Ryan – to update and expand his open access book R for Newspaper Data using a range of newly digitised newspapers and tools developed on Living with Machines.

Joanne Sheppard – to develop an interactive website aimed at rail enthusiasts using the StopsGB dataset.

Robert Sherman – to produce a visual poem, recreating the experience of rail travel using the StopsGB dataset.

Credits

Image Credits

4 Ruth Ahnert

7 Credit Léllé Demertzi

12 British railspace and buildings as predicted by a MapReader computer vision model.~30.5M patches from 16K nineteenth-century OS map sheets. (a) Predicted railspace; (b) predicted buildings; (c) and (d) predicted railspace (red) and buildings (black) in and around Middlesbrough and London, respectively. MapReader classifies information from large images or a set of images at a patch-level, as depicted in the figure insets. Map images courtesy of the National Library of Scotland (NLS).

14 Front page from Mitchell's Newspaper Press Directories, 63rd annual issue, 1908

16 Eight-Cylinder Revolving-Type Printing Machine in which we see humans physically within the mechanism: a mid-nineteenth-century machine made by American firm, R. Hoe & co., variously described as a 'monster' and whose components included 'turtles'. Source: Albert Sidney Bolles, *Industrial history of the United States*, 1878. No restrictions via Wikimedia Commons.

17 Ruth Ahnert, Emma Griffin, Mia Ridge, Giorgia Tolfo, Collaborative Historical Research in the Age of Big Data, Cambridge University Press, 2023

19 Credit David Lindsey

20 Credit Leeds City Museum

22 Credit Léllé Demertzi

Acknowledgements

With thanks to UKRIs' Strategic Priorities Fund for granting the funding for this project, and to the AHRC for administering the grant, and working with us in such a generous and collaborative fashion.

Thank you also to our advisory board (past and present) for their guidance and input 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, DCMS), David de Roure (University of Oxford), Jon Rowe (The Alan Turing Institute/ University of Birmingham), James Smithies (King's College London), Alan Sudlow (ex officio, AHRC), Melissa Terras (University of Edinburgh), and Jane Winters (School of Advanced Study).

OUR FUNDER

UKRI Arts and Humanities Research Council

OUR PARTNERS

The Alan Turing Institute

British Library

Queen Mary University of London

University of East Anglia

University of Exeter

University of Cambridge

King's College London

LIVING WITH MACHINES

livingwithmachines.ac.uk