DOI:
10.5334/johd.195

Document Version
Publisher's PDF, also known as Version of record

Journal of Open Humanities Data

Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project

DATA PAPER

MIA RIDGE
NILO PEDRAZZINI
MIGUEL VIEIRA
ARIANNA CIULA
BARBARA MCGILLIVRAY

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present the 'Language of Mechanisation' datasets with examples of re-use in visualisations and analysis. These reusable CSV files, published on the British Library's Research Repository, contain automatically-transcribed text from 19th century British newspaper articles. Volunteers on the Zooniverse crowdsourcing platform took part in tasks that asked 'How did the word *x* change over time and place?' They annotated articles with pre-selected meanings (senses) for the words *coach*, *car*, *trolley* and *bike*.

The datasets can support scholarship on a range of historical and linguistic research areas, including research on crowdsourcing and online volunteering behaviours, data processing and data visualisations methodologies.

CORRESPONDING AUTHOR:
Nilo Pedrazzini
The Alan Turing Institute, London, UK
npedrazzini@turing.ac.uk

KEYWORDS:
19th century British English; historical newspapers; historical semantics; transport history; crowdsourcing; data visualisation

TO CITE THIS ARTICLE:
Ridge, M., Pedrazzini, N., Vieira, M., Ciula, A., & McGillivray, B. (2024). Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project. *Journal of Open Humanities Data*, 10(1), 33, pp. 1–9. DOI: https://doi.org/10.5334/johd.195

(1) OVERVIEW

REPOSITORY LOCATION/DATA ACCESSIBILITY STATEMENT

British Library, London, UK. *Living with Machines* datasets are published on the British Library's Research Repository, https://bl.iro.bl.uk, in the Living with Machines collection: https://bl.iro.bl.uk/collections/1ecde964-4860-4f66-af33-e2b8ba487bf9.

The two datasets described are at:

- Language of Mechanisation: annotated historical newspaper articles https://doi.org/10.23636/5t9m-0g59
- OCR and crowdsourced annotations, Language of Mechanisation, JSON files https://doi.org/10.23636/z634-km37

CONTEXT

The *Living with Machines* project was a multidisciplinary research collaboration between the British Library and Alan Turing Institute with King's College London, the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London. It investigated how the introduction of machines into the workplace and landscape changed the lives of ordinary people in the long nineteenth century. Building on the British Library's work in the field,[1] crowdsourcing was a key component of the project, positioned both as a form of public engagement with our research and as a method for annotating text at scale.

Within *Living with Machines*, the work package (mini-project) *Language of Mechanisation*, led by Barbara McGillivray, analysed the changes in the English lexicon related to mechanisation. The effects of mechanisation in 19th-century Britain had strong repercussions on aspects of Late Modern English, particularly on its lexicon. During this dynamic moment in the history of English, its vocabulary underwent rapid changes in both written and spoken sources. New concepts and entities related to technological innovations, social changes, geographical expansion and contact with other languages led to the emergence of neologisms and loanwords, sometimes expressed through new meanings of existing words via metaphorical shifts, semantic change and innovation phenomena (cf. e.g., Gorlach, 1999; Kay & Allan, 2015; Sussman, 2009).

In 2022–23, we designed and ran a series of crowdsourcing tasks on the citizen science platform *Zooniverse*. We asked members of the public to close-read articles from 19th century newspapers that mentioned specific types of machines that we thought would yield interesting results on semantic change related to mechanisation in 19th-century English.

The resulting datasets also inspired some Notebooks, designed and developed by our colleagues from King's Digital Lab (KDL) (Miguel Vieira, Tiffany Ong and Arianna Ciula) in collaboration with Barbara McGillivray, Mia Ridge and Nilo Pedrazzini. The *Language of Mechanisation* Observable Notebook analyses words that changed meaning during the 19th century. Alongside studying temporal changes, we also explore instances where words varied in meaning across geographical regions or were employed differently in newspapers based on their political alignment.

(2) METHOD

The crowdsourcing tasks built on lessons from earlier explorations involving the public in creating data about changes in language on mechanisation. These tasks drew on a design process (Figure 1) adapted from library crowdsourcing projects.[2]

The aim of gathering crowdsourced annotations was to provide more finely grained and detailed information about the shifts that selected words underwent, while at the same time proving a validation test against which to compare the results of the algorithms.

1 https://www.bl.uk/projects/crowdsourcing-at-the-british-library (last accessed date: 14/10/2023).

2 The earlier tasks asked volunteers to transcribe text from articles about specific machines mentioned, and to match the 'machine' to senses from the Oxford English Dictionary (Ridge, 2020, 2022). Some results were visualised by Kalle Westerling for the project's exhibition, co-curated by Ridge (Leeds City Museum, July 2022–January 2023).

DOI: 10.5334/johd.195

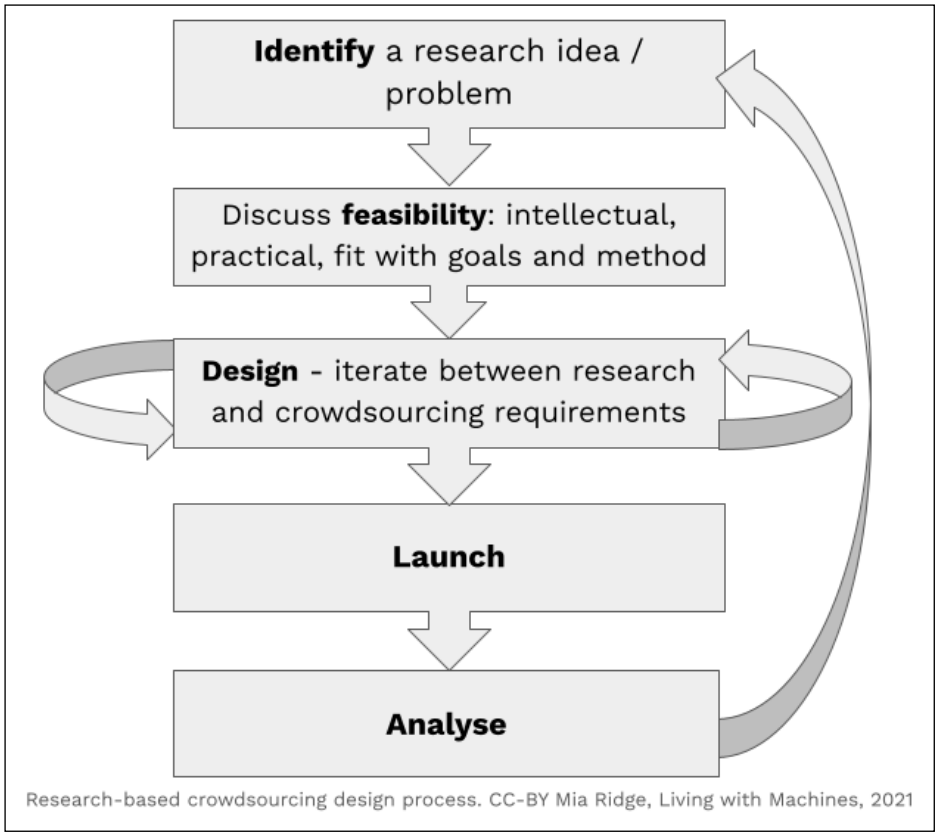

Figure 1 The process of designing crowdsourcing workflows for research tasks includes iterative discussions to ensure that tasks are both attractive to volunteers and produce valid research data.

SELECTING TARGET WORDS

The words *trolley*, *car*, *bike* and *coach* were selected as exemplars of the lexical field VEHICLE from a list of potential words that was obtained and subsequently narrowed down using the following criteria.

1. We used the Oxford English Dictionary (OED) Researcher API[3] to query the Historical Thesaurus of the OED (HTOED) and extracted all nouns belonging to the semantic category *Means of travel :: A conveyance* (HTOED category 03.10.03.02) or any of its sublevels. We limited the search to words that, according to the OED, had at least one new sense attested within the period of interest (1780–1920), as semantic change in the lexicon of mechanisation in 19th-century English was one of our overarching foci.
2. We sampled from different groups of nouns, classified by their likelihood of undergoing abrupt and radical semantic change in the 19th century. We used Pedrazzini and McGillivray's (2022a) decade-level diachronic word embeddings trained on 19th-century British newspapers digitised from the British Library's collection (1800–1919)[4] and followed the method described in Pedrazzini and McGillivray (2022b) to automatically detect significant changepoints. The approach involves setting a threshold[5] for what counts as a significant change: with a higher threshold, the algorithm becomes more conservative, requiring stronger evidence of a change before considering it significant. We took into account the results of using both a stricter (group 1) and a more liberal (group 2) threshold, to ensure that both highly likely and only potential semantic shifts were detected and kept separate. *Car* and *trolley* were selected from group 1, whereas *bike* and *coach* were selected from group 2.
3. During the task design process, some senses from the OED for the selected words were merged to obtain distinct definitions along clear-cut dimensions (*bike* and *trolley* along the non-motorised/motorised dimension; *car* and *coach* along both the non-motorised/motorised and road/railway dimensions).

3 https://languages.oup.com/research/oed-researcher-api/ (last accessed date: 14/12/2023).

4 https://www.bl.uk/collection-guides/newspapers (last accessed date: 3/6/2023).

5 This is the *penalty* hyperparameter for the Pelt algorithm (Killick, Fearnhead, & Eckley, 2012). As in Pedrazzini and McGillivray (2022b), we used the implementation of the Pelt algorithm from the Rapture library (Truong et al., 2020).

ARTICLE SELECTION AND SAMPLING STRATEGY

To sample the newspaper articles to be annotated via Zooniverse, we extracted occurrences of target words in newspaper articles via a full-text search of the newspaper collections provided by the British Newspaper Archive for the *Living with Machines* project. For each of the words we extracted at least 1,000 random occurrences and their context (i.e., the newspaper article in which they appear). Using the associated metadata, in our sampling we strived to obtain balance across different metadata values, focussing on having a diverse range of newspaper titles and years of publication represented. We achieved this with the help of Pandas (McKinney, 2010; The pandas development team, 2023) `sample` method using those two variables as values of the parameter weights. This step was first performed on all digitised articles with an estimated optical character recognition (OCR) quality of >0.90 only. If not enough representativeness among newspaper titles and year of publication could be achieved for a particular word with this method, the same step was first repeated on articles with an OCR quality between 0.70 and 0.90, and subsequently on those with an OCR quality of <0.70, until all newspaper titles were represented to some extent. Finally, we used Defoe (Filgueira et al., 2019) to extract the digitised images of each newspaper article listed in the subsample for each word and highlight the words of interest in the relevant article images.

TASK DESIGN

We designed four annotation tasks ('workflows') in the Zooniverse platform: *coach* (#23681), *car* (#23628), *trolley* (#23452), and *bike* (#23672). Content design activities included writing and testing text on the Zooniverse platform, including 'about' pages, workflow titles, descriptions, instructions and help texts (introductory tutorials, task help and 'Field Guides'), seen in part in Figure 2.

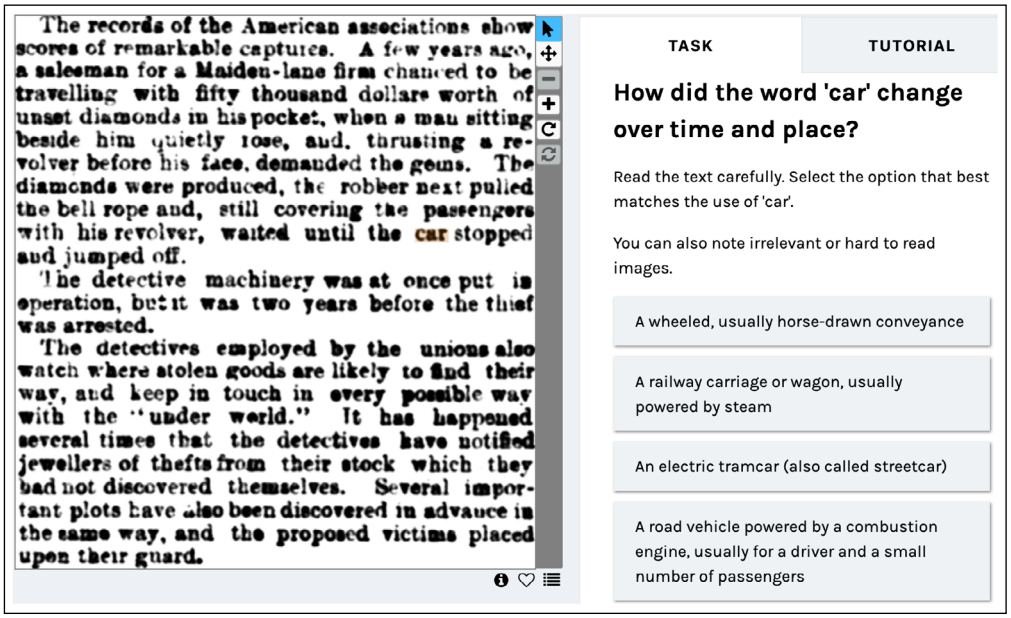

Figure 2 The Zooniverse task interface showing an extracted article and some annotation options for the word *car*.

Volunteers were recruited via social media, via the British Library's LibCrowds newsletter,[6] and via Zooniverse's live projects page.

QUALITY CONTROL FOR DATA ANALYSIS

Each article related to a word was annotated by multiple volunteers. To support the data analysis in the interactive Observable Notebook created with KDL, we calculated annotation agreement for each article. The average agreement was 83%. For subsequent analysis, we selected only the annotations that met a minimum agreement threshold of 65% (i.e., 2 out of 3 annotators agreed); this parameter can be configured by the user.

6 https://us11.campaign-archive.com/home/?u=08e409d3d85876a17ac4c1d09&id=e52e46328f (last accessed date: 14/12/2023).

(3) DATASET DESCRIPTION

REPOSITORY NAME

British Library

CODE REPOSITORIES

Two code repositories were used to process raw Zooniverse data to create the datasets discussed here. A Jupyter Notebook that anonymises usernames and IDs with a stable generated identifier was updated to process data for deposit (Westerling et al., 2023). These CSV files combine Zooniverse data about volunteer actions ('classifications') with information about the images ('subjects') uploaded to Zooniverse.

A second set of Notebooks created by KDL[7] summarises workflow activity, links it to newspaper metadata[8] (Figure 3) and exports in JSON format (without personally identifying data); exported data is then used in an Observable Notebook, discussed below.

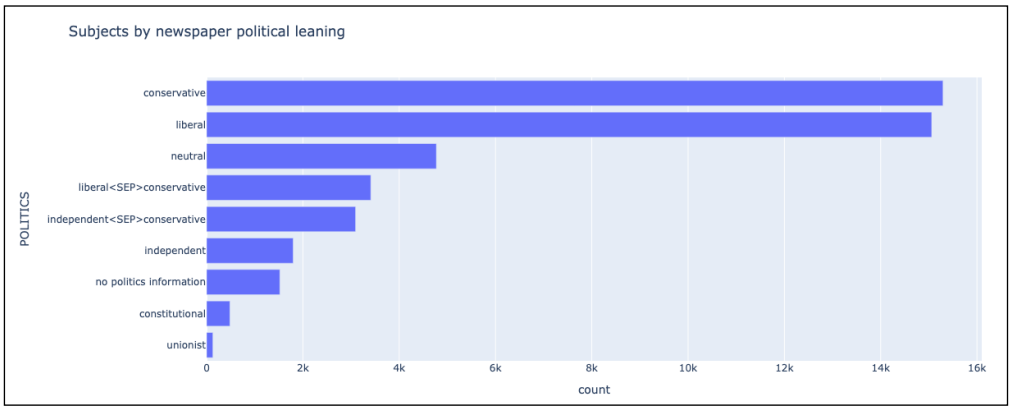

Figure 3 A graph showing the number of images from papers with different political leanings (KDL Explore Notebook).[9]

OBJECT NAMES (SEE TABLE 1)

Table 1 Dataset description.

CSV FILES, PROCESSED BY WORKFLOW						
WORKFLOW QUESTION	**ID**	**DATASET URL**	**OBJECT NAME**	**DATES ACTIVE**	**NUMBER OF ANNOTATIONS**	**NUMBER OF VOLUNTEERS**
How did the word 'trolley' change over time and place?	#23452	https://bl.iro.bl.uk/downloads/3f3d5097-b7b4-44e8-ab0d-dd45eca72721	combined-23452-trolley-classifications.csv	2023-02-06 to 2023-03-19	3,534 contributions on 1,006 completed subjects	138 volunteers (62 registered, 76 unregistered)
How did the word 'car' change over time and place?	#23628	https://bl.iro.bl.uk/downloads/ecd02aee-8315-4de2-8f54-d58dfb430718	combined-23628-car-classifications.csv	2023-02-27 to 2023-05-16	6,383 contributions on 1,993 completed subjects.	217 volunteers (135 registered, 82 unregistered)
How did the word 'coach' change over time and place?	#23681	https://bl.iro.bl.uk/downloads/330f129e-3599-455c-b91b-0dcdd7f053fb	combined-23681-coach-classifications.csv	2023-03-23 to 2023-05-15	6,583 contributions on 1,999 completed subjects	187 volunteers (104 registered, 83 unregistered)
Bicycle or motorcycle?	#23672	https://bl.iro.bl.uk/downloads/eba2e49d-6760-4991-a475-50ac3d74e5a8	combined-23672-bike-classifications.csv	2023-03-07 to 2023-03-19	7,754 contributions on 2,516 completed subjects	125 volunteers (69 registered, 56 unregistered)
README Language_of_Mechanisation_README___Data_Card_Mia_Ridge.docx https://bl.iro.bl.uk/downloads/3e1ed30d-5baa-457d-b894-8148c89637ef						

OBSERVABLE FILES, ORGANISED FOR VISUALISATION	
FILES	**DATASET URL**
annotations.json	https://bl.iro.bl.uk/downloads/2fff1f51-b16c-4363-a664-9f0f6e9d277b
county_region_mapping.csv	https://bl.iro.bl.uk/downloads/0b8fe042-5a0c-4f0c-9ada-66d422b6e0fd

(Contd.)

7 https://zenodo.org/records/10401205 (last accessed date: 14/12/2023).

8 https://doi.org/10.23636/pbq5-9k28 (last accessed date: 14/12/2023).

9 https://github.com/kingsdigitallab/lwm-davizct/blob/main/notebooks/Explore.ipynb (last accessed date: 14/12/2023).

OBSERVABLE FILES, ORGANISED FOR VISUALISATION	
FILES	**DATASET URL**
date.json	https://bl.iro.bl.uk/downloads/14b89f18-4a7f-4996-933d-038f0edce100
newspapers.json	https://bl.iro.bl.uk/downloads/980baa2e-6b52-4d7e-ab56-404515fcb366
participants.json	https://bl.iro.bl.uk/downloads/030916cf-4ae7-452a-9c69-cd18b7fd78bb
projects.json	https://bl.iro.bl.uk/downloads/a0774e7b-b883-424a-9c7b-687976038e0a
subjects.json	https://bl.iro.bl.uk/downloads/acb96b91-0ada-4597-8ac3-09c72240fef8
subjects_image.json	https://bl.iro.bl.uk/downloads/3d62f387-43c7-4baa-b86b-0c3abc732210
workflows.json	https://bl.iro.bl.uk/downloads/1edd9317-f11a-4c0e-8d5b-33baf19b6deb
README README.md https://bl.iro.bl.uk/downloads/1eb4e4f9-0a8e-4b85-a8d4-40b34947a44d	

FORMAT NAMES AND VERSIONS

Our dataset contains two sets of files:

1. One CSV file for each workflow (*trolley*, *car*, *coach* and *bike*), designed for general use and including the crowdsourced annotations, automatically-transcribed text of the article excerpt from digitised newspapers, with metadata related to both the annotated newspaper article (newspaper title, place of publication, article date, region of interest (ROI) within the image of the page) and its annotation on the Zooniverse platform (name of the workflow, time of annotation, number of annotations per item, etc.). A detailed README file explains each field.[10]
2. JSON files designed for the visualisations in KDL's Observable Notebook[11] and information about the content and purpose of each file in a README.[12]

CREATION DATES

2023-02 to 2023-05-15

DATASET CREATORS

The original writers, editors and publishers of 19th century newspapers on the British Newspaper Archive created the articles included in this data.

5,587 Zooniverse volunteers contributed to *Living with Machines* datasets overall.

LANGUAGE

English.

LICENSE

CC-BY 4.0 International

PUBLICATION DATE

2023

(4) REUSE POTENTIAL

The datasets described here are the first annotated sense datasets for historical English and will be relevant to historical linguists interested in the evolution of the English lexicon in the 19th century. The transcribed newspaper articles may be of interest to historians and others studying mechanisation in British society. Zooniverse metadata enables research into

10 https://bl.iro.bl.uk/downloads/3e1ed30d-5baa-457d-b894-8148c89637ef.

11 https://observablehq.com/@jmiguelv/language-of-mechanisation (last accessed date: 14/12/2023).

12 https://bl.iro.bl.uk/downloads/1eb4e4f9-0a8e-4b85-a8d4-40b34947a44d.

crowdsourcing and online volunteering patterns of behaviour. Finally, the datasets might be relevant for Research Software Engineers and Research Software UI/UX Designers as inspiration for processing and visualising annotated texts.

As an example of the reuse potential of our dataset in historical linguistics and historical research, Figure 4, generated by our Observable Notebook, shows how the datasets can be processed, visualised and analysed to explore how the target words changed meaning during the 19th century.

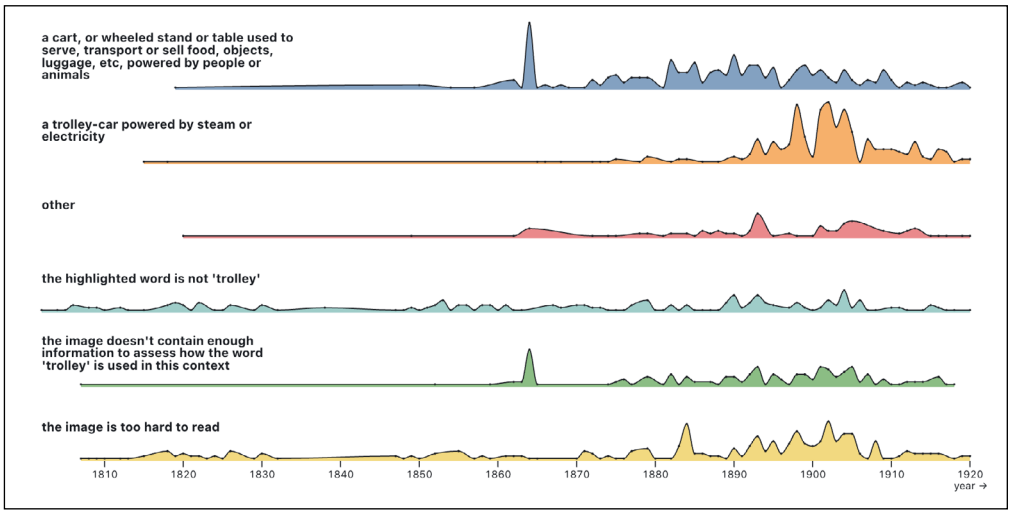

Figure 4 Raw total count of annotations for each word meaning (trolley as cart vs. trolley-car vs. trolley as other meaning) and other classifications.

ACKNOWLEDGEMENTS

Many members of the Living with Machines project contributed to related work over time, but we are particularly grateful to Kalle Westerling, Giorgia Tolfo and Claire Austin; FindMyPast, and the over 5,500 volunteers who contributed to our Zooniverse projects.

Tiffany Ong, KDL senior Research Software UI/UX Designer, contributed to the design (including user research and usability testing) and description of the dynamic Notebooks visualisations.

Feedback on the usability of the Observable visualisations was provided by British Library staff members. Pam Mellen and Shalen Fu (KDL Lab Manager and Project Manager at the time) contributed to the administration of the project for the design of the interactive Notebooks.

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including from the National Science Foundation, NASA, the Institute of Museum and Library Services, UK Research and Innovation (UKRI) , Google (Global Impact Award), and the Alfred P. Sloan Foundation.

FUNDING INFORMATION

Living with Machines, funded by the UKRI Strategic Priority Fund, was a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library and Cambridge, King's College London, East Anglia, Exeter, and Queen Mary University of London (Grant Reference: AH/S01179X/1).

Newspaper data was provided by Findmypast Limited from the British Newspaper Archive, a partnership between the British Library and Findmypast. See https://www.britishnewspaperarchive.co.uk/ for more details.

The writing and publication of this paper was in part supported by the Ecosystem Leadership Award (EPSRC Grant EP/X03870X/1) and The Alan Turing Institute, particularly the Turing Research Fellowship scheme, and by *Data/Culture: Building sustainable communities around arts and humanities datasets and tools*, a collaborative pilot project between The Alan Turing Institute, Queen Mary University London, Lancaster University, and the Complexity Science Hub, funded by the Arts and Humanities Research Council (AHRC; Grant Ref: AH/Y00745X/1) and part of UKRI.

COMPETING INTERESTS

McGillivray is editor-in-chief of the Journal of Open Humanities Data but did not take part in the editorial process or decisions pertaining to this manuscript.

AUTHOR CONTRIBUTIONS

MR: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition; Project administration; Resources; Supervision.

NP: Methodology, Writing – original draft, Writing – review & editing.

MV: Methodology; Data curation; Software; Validation; Visualization; Writing – review & editing.

AC: Methodology; Visualization; Project administration; Writing – review & editing.

BMcG: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, supervision, Funding acquisition.

AUTHOR AFFILIATIONS

Mia Ridge orcid.org/0000-0003-3733-8120
Digital Research, British Library, London, UK

Nilo Pedrazzini orcid.org/0000-0003-3757-2961
The Alan Turing Institute, London, UK

Miguel Vieira orcid.org/0009-0005-4189-8739
King's Digital Lab, King's College London, London, UK

Arianna Ciula orcid.org/0000-0003-4247-1073
King's Digital Lab, King's College London, London, UK

Barbara McGillivray orcid.org/0000-0003-3426-8200
Department of Digital Humanities, King's College London, London, UK;
The Alan Turing Institute, London, UK

REFERENCES

Filgueira, R., Jackson, M., Terras, M., Beavan, D., Roubickov, A., Hobson, T., Coll Ardanuy, M., Colavizza, G., Krause, A., Hetherington, J., Hauswedell, T., Nyhan, J., & **Ahnert, R.** (2019). defoe: A Spark-Based Toolbox for Analysing Digital Historical Textual Data. *2019 5th International Conference on eScience*, 235–242. DOI: https://doi.org/10.1109/eScience.2019.00033

Gorlach, M. (1999). English in Nineteenth-Century England. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511627828

Kay, C., & **Allan, K. L.** (2015). Historical Semantics. Edinburgh: Edinburgh University Press. DOI: https://doi.org/10.1515/9780748644797

Killick, R., Fearnhead, P., & **Eckley, I. A.** (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590–1598. DOI: https://doi.org/10.1080/01621459.2012.737745

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, Austin, 56–61. DOI: https://doi.org/10.25080/Majora-92bf1922-00a

Pedrazzini, N., & **McGillivray, B.** (2022a). *Diachronic word embeddings from 19th-century newspapers digitised by the British Library (1800–1919)*. [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.7181682

Pedrazzini, N., & **McGillivray, B.** (2022b). Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers. *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, Taipei, 85–95. https://aclanthology.org/2022.nlp4dh-1.12.

Ridge, M. (27 October 2020). What Is a "Machine" Anyway? Help Us Describe Them. *Living with Machines*. https://livingwithmachines.ac.uk/what-is-a-machine-anyway-help-us-find-out/

Ridge, M. (22 March 2022). Zooniverse Activity, March 2022. *Living with Machines*. https://livingwithmachines.ac.uk/zooniverse-activity-march-2022/

Sussman, H. (2009). *Victorian Technology: Invention, Innovation, and the Rise of the Machine*. London: Bloomsbury Publishing. DOI: https://doi.org/10.5040/9798216032052

The pandas development team. (2023). *pandas-dev/pandas: Pandas* (V2.1.3). [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.10107975

Truong, C., Oudre, L., & **Vayatis, N.** (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*(107299). DOI: https://doi.org/10.1016/j.sigpro.2019.107299

Westerling, K., Tolfo, G., Pedrazzini, N., & **Ridge, M.** (15 December 2023). *Notebook: Prepare Zooniverse Data for Analysis and Deposit* (V0.1.1-beta). [Data set]. British Library. DOI: https://doi.org/10.5281/ZENODO.10392954

TO CITE THIS ARTICLE:
Ridge, M., Pedrazzini, N., Vieira, M., Ciula, A., & McGillivray, B. (2024). Language of Mechanisation Crowdsourcing Datasets from the Living with Machines Project. *Journal of Open Humanities Data*, 10(1), 33, pp. 1–9. DOI: https://doi.org/10.5334/johd.195

Submitted: 03 January 2024
Accepted: 22 March 2024
Published: 29 April 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.