Metadata Enrichment in the Living with Machines Project: User-focused Collaborative Database Development in a Digital Humanities Context

Westerling, Kalle
kalle.westerling@bl.uk
British Library/The Alan Turing Institute, United Kingdom

Beavan, David
dbeavan@turing.ac.uk
The Alan Turing Institute

Beelen, Kaspar
kbeelen@turing.ac.uk
The Alan Turing Institute

Coll Ardanuy, Mariona
mcollardanuy@gmail.com
The Alan Turing Institute

Hobson, Timothy
thobson@turing.ac.uk
The Alan Turing Institute

Last, Christina
lastc@mit.edu
The Alan Turing Institute

Pedrazzini, Nilo
npedrazzini@turing.ac.uk
The Alan Turing Institute

Reese, Griffith
grees@turing.ac.uk
The Alan Turing Institute

Luke, Hare
lhare@turing.ac.uk
The Alan Turing Institute

Living with Machines (LwM) brings historians, data scientists, geographers, computational linguists, and curators together to explore the impact of technology on the lives of ordinary people during the Industrial Revolution. The project harnesses the combined power of massive digitised historical collections and computational tools to examine ways in which technology altered the fabric of human existence on a hitherto unprecedented scale. We use 100+ billion words (20+ terabytes) of 19th century newspaper textual data, and related metadata jointly digitised by the British Library, FindMyPast and Jisc. The ability to spatially and temporally query across this data at scale is crucial. The project shares these challenges with other projects that rely on similar content, such as Oceanic Exchanges (Beals / Bell 2020), Impresso (Romanello et al. 2020), and NewsEye (Jean-Caurant / Doucet 2020). Unlike many of those projects, we did not intend to build a search interface. Rather, we collaboratively developed infrastructure to analyse data in an efficient, reproducible way to answer research questions.

We collaborated with LwM's researchers to identify use cases for the newspaper metadata and saw the need for a flexible digital research infrastructure to improve our pipeline and provide a generalizable resource for other researchers working on historical newspaper collections. This poster targets digital humanities researchers working on large-scale historical newspaper collections with an aim to remove technical barriers to make humanities research on large, heterogeneous, and complex newspaper data collections quicker, easier and more reproducible.

With a team fluent in Python and where Python is the predominant language across packages we used and built, we chose to work with Django: a Python package primarily used for fullstack web development. We chose it because of its well documented and intuitive implementation of an Object Relational Mapper (ORM) to provide an Python API layer between the database backend and the "frontend" where researchers would use Jupyter notebooks for accessing the database through the Python layer. The GeoDjango features also offered a straightforward method to store and spatially query our geocoded data. Our work was structured in a five-step parallel and iterative process: Designing, Developing, Deploying, and, finally, Documentation and Dissemination (Ahnert et al. 2023). Consequently, this poster describes the radically collaborative workflow of building, packaging, and disseminating the database infrastructure, the software layer, and the populated database.

In the design phase, the 15–20 person team met in six hybrid full-day workshops discussing early prototypes developed by research software engineers and research data scientists. These workshops aimed to utilise the team's multidisciplinary expertise and collaborative project community to combine structured data from historical documents, including press directories and gazetteers, beyond the newspapers (Beelen et al. 2022). They also aimed to reduce barriers caused by specialist language by establishing mutual vocabularies, clarifying data flows, and providing example code to demonstrate the database's possibilities for different levels of technological background for different team members.

During the development phase, the project's programmers utilised tools built by the project in a three-step process. The first step disambiguated different XML flavours for post-digitisation metadata, including Optical Character Recognition (OCR) and layout detection into standardised, minimal amount of metadata using the alto2txt tool (Smith et al. 2022). In the second step, the metadata was transformed into files for easier ingestion by Django. Finally, the backend Python/Django solution was built in its own repository (Beavan et al. 2023), facilitating containerisation in the next phase.

The deployment phase aimed to create a sustainable solution that could outlast the LwM's timeline by containerising the Django/Postgres backend database with PostGIS extension in Docker images. This allowed easy redeployment and use across different operating systems and architecture, making the data and

code open-access and open-source. A continuous integration, testing, linting, and deployment workflow was implemented, alongside deployment and user testing via the project's Microsoft Azure-based virtual machines.

The documentation and dissemination phase was integral throughout the process, with an assigned lead maintaining a user-centred focus by creating tutorials, Jupyter notebooks, and detailed instructions during package design, development, and deployment. The dissemination phase aimed to build a user community by following good documentation practices and ideas for building engaged potential user communities with clear examples and tutorials. It also aimed to build contributor communities, with the software available in a GitHub repository to attract potential open-source collaborators (i.e. other Research Software Engineers, etc.) who would get appropriate credit for their code contributions. Here, we followed industry standards in providing transparent instructions for potential maintainers and contributors to the codebase (Trinkenreich et al. 2020; Katz et al. 2019; Pinto / Steinmacher / Gerosa 2016; Dias et al. 2021).

Bibliography

Ahnert, Ruth / **Griffin, Emma** / **Tolfo, Giorgia** / **Ridge, Mia** (2023): *Collaborative Historical Research in the Age of Big Data: Lessons from an interdisciplinary project*. Cambridge: Cambridge University Press.

Beals, H. M. / **Bell, Emily** (2020): *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough University. https://figshare.com/articles/online_resource/The_Atlas_of_Digitised_Newspapers_and_Metadata_Reports_from_Oceanic_Exchanges/11560059/2 [28.04.2023].

Beavan, David / **Last, Christina** / **Westerling, Kalle** / **Rees, Griffith** / **Pedrazzini, Nilo** / **Hobson, Timothy** / **Coll Ardanuy, Mariona** (2023): *Living with Machines Database: lwmdb*. https://github.com/living-with-machines/lwmdb.

Beelen, Kaspar / **Lawrence, Jon** / **Wilson, Daniel C S** / **Beavan, David** (2022): "Bias and representativeness in digitized newspaper collections: Introducing the environmental scan", in: *Digital Scholarship in the Humanities*. 10.1093/llc/fqac037.

Dias, Edson / **Meirelles, Paulo** / **Castor, Fernando** / **Steinmacher, Igor** / **Wiese, Igor** / **Pinto, Gustavo** (2021): *What Makes a Great Maintainer of Open Source Projects?*. in: *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. 982–994.

Jean-Caurant, Axel / **Doucet, Antoine** (2020): *Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform*. in: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. New York, NY, USA: Association for Computing Machinery. 531–532. https://dl.acm.org/doi/10.1145/3383583.3398627 [28.04.2023].

Katz, Daniel S. et al. (2019): "Community Organizations: Changing the Culture in Which Research Software Is Developed and Sustained", in: *Computing in Science & Engineering* 21 (2): 8–24. 10.1109/MCSE.2018.2883051.

Pinto, Gustavo / **Steinmacher, Igor** / **Gerosa, Marco Aurélio** (2016): *More Common Than You Think: An In-depth Study of Casual Contributors*, in: *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. 112–123.

Romanello, Matteo / **Ehrmann, Maud** / **Clematide, Simon** / **Guido, Daniele** (eds.) (2020): *The impresso system architecture in a nutshell*. The Hague: EuropeanaTech Insights.

Smith, Andy / **Jackson, Michael** / **Van Strien, Daniel** / **Beavan, David** / **Rees, Griffith** / **France, Lydia** / **Ryan, Yann** / **Nanni, Federico** (2022): Living-with-machines/alto2txt. Zenodo. https://zenodo.org/record/7378349 [10.01.2023].

Trinkenreich, Bianca / **Guizani, Mariam** / **Wiese, Igor** / **Sarma, Anita** / **Steinmacher, Igor** (2020): "Hidden Figures: Roles and Pathways of Successful OSS Contributors", in: *Proceedings of the ACM on Human-Computer Interaction 4 (CSCW2)*: 180:1-180:22. 10.1145/3415251.