Introduction

My project's main goal, as outlined in the initial proposal, was to develop an interactive, open-source web app that visualizes data from the Press Directories dataset alongside historical general election results. In this report, I will delve into the challenges encountered, the interesting findings uncovered, and the potential avenues for further research and improvement.

The project

The main final output of this project is an interactive web app that allows users to explore the political affiliations of newspapers distributed in UK counties to easily compare them with neighbouring counties, as well as view seat allocation results from the nearest general election using data coming from the CLEA project. The code for the web app, as well as the code for cleaning the datasets is available on this GitHub repository: github.com/Living-with-machines/VisualisingPressDirectories

Beyond historical exploration, the project's aim was also to add value by aligning the Press Directories dataset, which spans a period of suffrage expansion, with cleaned election data, enhancing its original environmental analysis purpose. To that end, other outputs of the project consist in cleaning processes and data replacements that extend the Living with Machines' work on the Mitchell Press Directories dataset.

Challenges faced

As with any project, the initial design and analysis process for the proposal had to come to terms with the realities of available data, which did not, however, impose a significant departure from the original plan.

Mapping the Press Directories

One of the first challenges I encountered was sourcing pertinent mapping data to correlate with the Press Directories dataset, which maps each newspaper to a *district*, which is then associated with a *county*. I decided to use the *counties*, which mostly matched with the United Kingdom's historical counties names contained in the map I used for the visualisation, based on data coming from the Historic County Borders Project.

Presenting data clearly and effectively

To achieve a clear and readable representation of the datasets, I implemented two main strategies:

- Grouping political leanings: I categorised political leanings and introduced a comprehensive "other" category for county-specific classifications, ensuring a more coherent mapping while maintaining the possibility of representing most ideological majorities at a county level.
- Unified colour scheme: Aiming for clarity and consistency, I strived to employ a unified colour scheme throughout the visualizations to facilitate understanding, trying to match as best as I could newspapers' politics to existing parties, with the obvious limitations due to imperfect overlap of political ideologies and party representation.

Pre-processing technical workload

Managing the technical aspects of this project posed another challenge. The initial proposal used Python on the back end to reshape datasets live and provide data visualisations directly from the original sources. Given the complexity of pre-processing tasks, especially regarding the electoral datasets, this proved impossible, so I switched to a hybrid approach involving both Python-based preprocessing and live visualisation with JavaScript based on JSON exports.

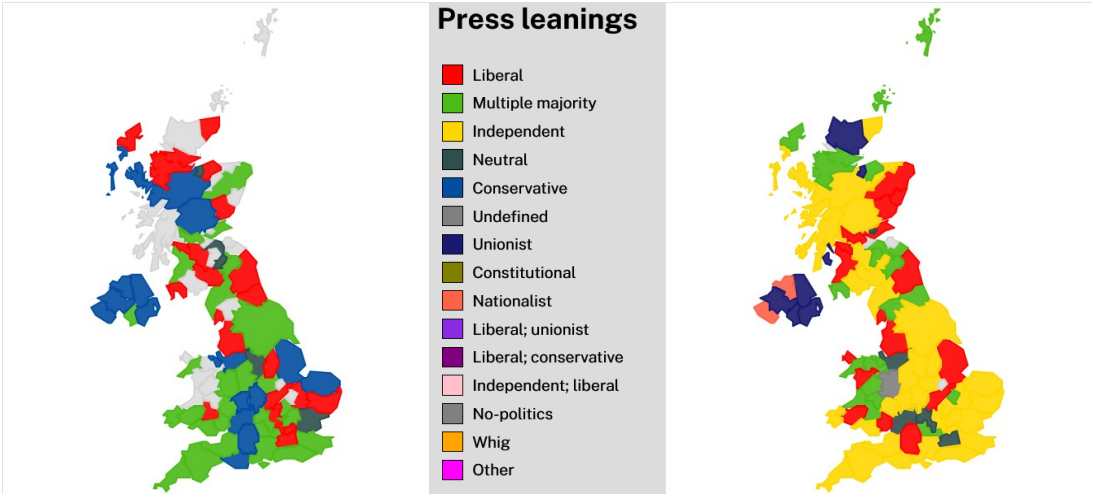
Interesting findings

While the main goal of the project was the web app, of course it allowed me to explore the Press Directories dataset in detail and find fascinating insights which I partially summarised below:

Changing political landscape

This was a two-fold change, which added to its interest: most notably, the electoral landscape witnessed a certain level of consistency, with county-level majorities keeping in line for most of the time period, except for the Labour party gradually replacing the

Liberal party in the early 20th century. On the other hand, during those same decades there was a remarkable proliferation of independent newspapers, resulting in independent majorities across the country, as shown in the figures illustrating the change between 1846 and 1922 which display the most common press leaning for each county.



Enriching datasets

This project gave me an increased appreciation for the value of datasets such as the CLEA one for political results and the historic counties project for mapping. By bridging these datasets with others created by the Living with Machines project (and other digital humanities endeavours) we have the potential to create a richer and more comprehensive resource for future research.

Data cleanup and integration

A pivotal aspect of this project involved cleaning up datasets and establishing links with other relevant datasets. In the course of this process, I realised the potential of fuzzy match techniques with human supervision, which greatly accelerated my capacity to find typos and over specific county labels.

Helpful techniques notwithstanding, the data cleanup was still challenging and time consuming, so I hope and believe my effort will enhance the overall quality of the resources and simplify the work of others. If you want to know more about the process, you can find more information in the Jupyter notebooks available on the projects' repository.

Limitations and opportunities

The short turnaround schedule for the project and its well-defined scope provided a comfortable environment in which to produce the final output while keeping track of possible improvements and avenues to grow. To that end, I created this overview of elements that I believe would be useful for those who might wish to expand on this project or create a similar one.

1. Exploring Press categorisation further: while all political leanings listed by Mitchell would benefit from further scrutiny, this applies most importantly to the

label and concept of "independent" press and its significance, especially given its growing presence throughout the time period under consideration.
2. Enhancing electoral data quality: outside the scope of the Press Directories dataset, while the electoral data available through the CLEA project is invaluable, it is not without its flaws, particularly in terms of constituency names, where multiple typos and unclear spelling can be found. Efforts to rectify these issues would be beneficial to this and many other projects.
3. Addressing boundary data challenges: obtaining accurate boundary data proved to be an impossible task, with relevant datasets often being private. The fuzzy matching and geolocation method employed to make up for this also had its limitations, compounded by typos in electoral datasets.
4. Colour schemes: finally, and most pragmatically, the choice of colour schemes I made for political leanings does not always align perfectly with political parties, and there is certainly room for debate and refinement in this regard, a debate that would benefit from expert supervision from people more familiar with historical political persuasions and parties.

Final remarks

In my project, I aimed to achieve two main goals. Firstly, I wanted to make the valuable Press Directories dataset more accessible to scholars by creating a user-friendly tool. This tool serves as an immediate resource, especially for those who might not typically use large datasets. Additionally, I hope it will attract individuals interested in the topic who may not be aware of the resources offered by the Living with Machines project.

Secondly, I valued the opportunity to enhance and expand upon these resources as a part of open source and open science principles. This included proposing corrections to the Press Directories dataset and the CLEA one, which I mentioned in this report and proposed to the repository's maintainers.

Lastly, I want to express my gratitude for the project's support and innovation. In my experience, the project staff has been attentive and helpful. If I were to suggest an improvement, it would involve introducing a user research approach where beta versions of tools and outputs intended for public use could be shared with the Institute's audience. This would enable iterative improvements while encouraging community participation.