The Past, Present and Future of Digital Scholarship with Newspaper Collections[⋆]

Mia Ridge[1], Giovanni Colavizza[2], Laurel Brake[3], Maud Ehrmann[4], Jean-Philippe Moreux[5], and Andrew Prescott[6]

[1] British Library, UK mia.ridge@bl.uk
[2] The Alan Turing Institute, UK gcolavizza@turing.ac.uk
[3] Birkbeck, University of London, UK l.brake@bbk.ac.uk
[4] Digital Humanities Laboratory, EPFL, Switzerland maud.ehrmann@epfl.ch
[5] Bibliothèque nationale de France, France jean-philippe.moreux@bnf.fr
[6] University of Glasgow, UK andrew.prescott@glasgow.ac.uk

Abstract. Historical newspapers are of interest to many humanities scholars as sources of information and language closely tied to a particular time, social context and place. Digitised newspapers are also of interest to many data-driven researchers who seek large bodies of text on which they can try new methods and tools. Recently, large consortia projects applying data science and computational methods to historical newspapers at scale have emerged, including NewsEye, *impresso*, Oceanic Exchanges and Living with Machines.
This multi-paper panel draws on the work of a range of interdisciplinary newspaper-based digital humanities and/or data science projects, alongside ’provocations’ from two senior scholars who will provide context for current ambitions. As a unique opportunity for stakeholders to engage in dialogue, for the DH2019 community to ask their own questions of newspaper-based projects, and for researchers to map methodological similarities between projects, it aims to have a significant impact on the field.

[⋆] Panel presented at the 2019 Digital Humanities Conference, Utrecht, July 2019 https://dev.clariah.nl/files/dh2019/boa/0377.html

1 Panel Overview

Historical newspapers are of interest to many humanities scholars, valued as sources of information and language closely tied to a particular time, social context and place. Following library and commercial microfilming and, more recently, digitisation projects, newspapers have been an accessible and valued source for researchers. The ability to use keyword searches through more data than ever before via digitised newspapers has transformed the work of researchers (as discussed by others including Putnam [2016], Bingham [2010]).

Digitised historic newspapers are also of interest to many researchers who seek large bodies of relatively easily computationally-transcribed text on which they can try new methods and tools. Intensive digitisation over the past two decades has seen smaller-scale or repository-focused projects flourish in the Anglophone and European world Holley [2009], King [2005], Neudecker et al. [2014]. However, just as earlier scholarship was potentially over-reliant on The Times of London and other metropolitan dailies, this has been replicated and reinforced by digitisation projects (for a Canadian example, see Milligan [2013]).

In the last years, several large consortia projects proposing to apply data science and computational methods to historical newspapers at scale have emerged, including NewsEye, *impresso*, Oceanic Exchanges and Living with Machines. This panel has been convened by some consortia members to cast a critical view on past and ongoing digital scholarship with newspapers collections, and to inform its future.

Digitisation can involve both complexities and simplifications. Knowledge about the imperfections of digitisation, cataloguing, corpus construction, text transcription and mining is rarely shared outside cultural institutions or projects. How can these imperfections and absences be made visible to users of digital repositories? Furthermore, how does the over-representation of some aspects of society through the successive winnowing and remediation of potential sources - from creation to collection, microfilming, preservation, licensing and digitisation - affect scholarship based on digitised newspapers. How can computational methods address some of these issues?

The panel proposes the following format: short papers will be delivered by existing projects working on large collections of historical newspapers, presenting their vision and results to date. Each project is at different stages of development and will discuss their choice to work with newspapers, and reflect on what have they learnt to date on practical, methodological and user-focused aspects of this digital humanities work. The panel is additionally an opportunity to consider important questions of interoperability and legacy beyond the life of the project. Two further papers will follow, given by scholars with significant experience using these collections for research, in order to provide the panel with critical reflections. The floor will then open for debate and discussion.

This panel is a unique opportunity to bring senior scholars with a long perspective on the uses of newspapers in scholarship together with projects at formative stages. More broadly, convening this panel is an opportunity for the DH2019 community to ask their own questions of newspaper-based projects, and for researchers to map methodological similarities between projects. Our hope is that this panel will foster a community of practice around the topic and encourage discussions of the methodological and pedagogical implications of digital scholarship with newspapers.

2 Living with Machines

***Paper authors**: Giovanni Colavizza[1], Mia Ridge[2] with Ruth Ahnert[3], Claire Austin[2], David Beavan[1], Kaspar Beelens[1], Mariona Coll Ardanuy[1], Adam Farquhar[2], Emma Griffin[4], James Hetherington[1], Jon Lawrence[5], Katie McDonough[1], Barbara McGillivray[6], André Piza[1], Daniel van Strien[2], Giorgia Tolfo[2], Alan Wilson[1], Daniel Wilson[1].*

***Author institutes**: 1 Alan Turing Institute; 2 British Library; 3 Queen Mary University of London; 4 University of East Anglia; 5 University of Exeter; 6 Alan Turing Institute/University of Cambridge.*

Living with Machines is a five-year interdisciplinary research project, whose ambition is to blend data science with historical enquiry to study the human impact of the industrial revolution. Set to be one of the biggest and most ambitious digital humanities research initiatives ever to launch in the UK, Living with Machines is developing a large-scale infrastructure to perform data analyses on a variety of historical sources, and in so doing provide vital insights into the debates and discussions taking place in response to today's digital industrial revolution.

Seeking to make the most of a self-described 'radical collaboration', the project will iteratively develop research questions as computational linguists, historians, library curators and data scientists work on a shared corpus of digitised newspapers, books and biographical data (census, birth, death, marriage, etc. records). For example, in the process of answering historical research questions, the project could take advantage of access to expertise in computational linguistics to overcome issues with choosing unambiguous and temporally stable keywords for analysis, previously reported by others Lansdall-Welfare et al. [2017]. A key methodological objective of the project is to 'translate' history research questions into data models, in order to inspect and integrate them into historical narratives. In order to enable this process, a digital infrastructure is being collaboratively designed and developed, whose purpose is to marshal and interlink a variety of historical datasets, including newspapers, and allow for historians and data scientists to engage with them.

In this paper we will present our vision for Living with Machines, focusing on how we plan to approach it, and the ways in which digital infrastructure enables this multidisciplinary exchange. We will also showcase preliminary results from the different research 'laboratories', and detail the historical sources we plan to use within the project.

3 *impresso* - Media Monitoring of the Past

Paper authors*: Maud Ehrmann[1], Matteo Romanello[1], Frédéric Kaplan[1], Marten Düring[2], Estelle Bunout[3], Daniele Guido[2], Paul Schroeder[2], Thijs van Beek[2], Andreas Fickers[2], Simon CLematide[3], Phillip Ströbel[3], Martin Volk[3].*

Author institutes*: 1 École polytechnique fédérale de Lausanne, DHLAB; 2 University of Luxembourg, C^2DH; 3 University of Zurich, Institute of Computational Linguistics.*

Historical newspapers are mirrors of past societies. Published over centuries on a regular basis, they record wars and minor events, report on international, national and local matters, and document the day-to-day life; in a word, they keep track of the great and small history. They reflect the political, moral, and economic environments in which they were produced and they hold dense, continuous, and multi-level information which can help us understand how contemporaries experienced their present. This makes them invaluable primary sources for historians.

How can newspapers help understanding the past? How to explore them? Long held on library and archive shelving, newspapers are undergoing mass digitization. Millions of facsimiles, along with their machine-readable content acquired via optical character recognition (OCR), are becoming accessible via a variety of online portals. If this represents a major step forward in terms of preservation of and access to documents, much remains to be done in order to provide extensive and sophisticated access to the content of these digital resources. In this regard, we still face many challenges.

First, not all historical newspapers are digitized, and heterogeneous schemes of availability and accessibility lead to an opaque landscape of 'historical media silos'. Next, the quality of OCR outputs often makes subsequent automatic text processing difficult and unreliable. This content accessibility issue closely relates to the more fundamental – and promising – challenge of content exploitation and exploration: how to make sense of the vast amount of available unstructured text? To achieve this, we need to semantically enrich the contents of historical newspapers, i.e. to extract, process, and link the information they contain. Another challenge relates to data visualization and exploration which need to accompany enhanced text analysis capacities and comply with historical research imperatives. Finally, these challenges can only be met through the close interplay between computer sciences and history, an essential factor for enabling new and methodologically reflected digital history scholarship.

The interdisciplinary *impresso* project[7] aims to develop a methodologically-reflected technological framework to enable new ways of engaging with multilingual digital content of historical newspapers and new approaches to address historical questions. More precisely, the project applies text mining techniques to transform noisy and unstructured textual content into semantically indexed, structured, and linked data; develops innovative visualization interfaces to enable the seamless exploration of complex and vast amounts of historical data[8]; identifies needs on the side of historians which may also translate into new text mining applications and new ways to study history; and synergistically reflects on the usage of digital tools in historical sciences from a practical, methodological, and epistemological point of view.

We will try to answer the question ”how to build a historical media monitoring tool suite?”, by introducing the main objectives of *impresso* and present what was achieved as well as lessons learned so far. In particular, we will focus on how we manage, process and represent data, and how we operationalize interdisciplinary collaboration via interface co-design in order to welcome the scholarly use of the material.

Impresso is supported by the Swiss National Science Foundation under grant CR- SII5_173719.

4 Construire avec les usagers la numérisation des collections de périodiques

***Paper author**: Jean-Philippe Moreux.*
***Author institute**: Bibliothèque nationale de France.*

Les collections de périodiques anciens numérisés ont désormais un âge respectable, et le paysage des usages qu’elles auront contribué à dessiner a donc considérablement changé. Différentes politiques se sont succédé, depuis la numérisation d’une sélection restreinte de titres de presse quotidienne versée dans un portail documentaire, jusqu’aux programmes de numérisation de masse s’inscrivant dans un objectif de préservation. Les usages soutenus par ces politiques auront évolué de manière concomitante, de la recherche d’information aux pratiques plus récentes de fouille de données à finalité scientifique.

Une tension constante traverse cette période, celle de la documentation de ces politiques et de sa communication aux usagers. Si en 2007, il était aisé d’expliciter le choix de numériser Le Figaro, dix ans plus tard, ce titre est une ressource parmi 3 300 autres, toutes puisées dans le réservoir des 250 000 notices de périodiques conservées à la BnF. Cette trajectoire menant à une abondance de la ressource, participe cependant à alimenter le questionnement légitime de toutes

[7] *impresso* - Media Monitoring of the Past: `https://impresso-project.ch/`
[8] Impresso interface: `https://impresso-project.ch/`

les catégories d'usagers : quel est le contenu de la boîte noire que je suis en train d'interroger ? Et pour les utilisateurs relevant du champ des humanités numériques, ce questionnement est d'autant plus impératif qu'ils doivent prendre en compte les enjeux de description, de représentativité, de biais, etc. de leur corpus de travail.

Répondre à cette question, du point de vue des gestionnaires des bibliothèques numériques, n'est pas trivial, car elle agrège plusieurs dimensions et appelle de ce fait des actions multiples. Au niveau macroscopique, il s'agirait tout d'abord de décrire des choix documentaires (de numérisation) et de les situer dans un paysage plus vaste (national, européen). Les catalogues, les portails agrégateurs [9], participent de cet effort et aident les utilisateurs à découvrir et localiser les ressources à disposition. Mais ils ne dévoilent que rarement les politiques documentaires et techniques à l'œuvre. Des initiatives spécifiques au public des humanités numériques sont également nécessaires, par exemple le recensement ou la création de corpus de référence préparés pour des usages de fouille de données [10].

A une échelle inférieure, celle du périodique numérisé lui-même, l'utilisateur est le plus souvent confronté à la sécheresse d'une notice bibliographique, alors que les questions qu'il se pose sont nombreuses : contexte et historique de publication, exhaustivité (et manques), pratiques de numérisation, qualité du texte transcrit par l'OCR.

On peut imaginer que certains de ces défis devront être relevés par l'instauration d'une dynamique d'ouverture des données aux usagers et de coconstruction avec les usagers. Ainsi, enjamber les frontières, qu'elles soient nationales ou techniques (silos), se fera via des protocoles et API interopérables, autorisant la création de corpus ad hoc à des fins de recherche. Co-construire une politique de numérisation impliquera de donner aux utilisateurs la liberté d'émettre des suggestions de numérisation et aux équipes de recherche un rôle de partenaire actif des choix de numérisation. Enfin, enrichir la description des ressources numérisées, jusqu'au niveau élémentaire du numéro ou du fascicule, pourrait aussi s'appuyer sur l'intelligence collective des usagers.

5 Digital Editions of Serials and media historians: an overview

***Paper author**: Laurel Brake.*
***Author institute**: Birkbeck, University of London.*

The second reflective paper is from a 19th century media historian whose experience as a Principal Investigator includes working collaboratively to design and produce a curated online edition of runs of six 19th century serials across the century. Ncse, the Nineteenth-Century Serials Edition[11], includes a Chartist

[9] Voir par exemple presselocaleancienne.bnf.fr, www.zeitschriftendatenbank.de.
[10] `lab.kb.nl`, `api.bnf.fr`, `data.bl.uk`
[11] `https://ncse.ac.uk`

newspaper, a satirical illustrated paper, a professional journal, an early feminist monthly, a weekly news magazine, and an early theology journal. A decade old, ncse recently required upgrading, and this paper reflects this timely encounter with sustainability and how projects grow in the light of changing contexts.

Nsce defines its remit as ‘serials’, including ‘newspapers’ and magazines, but I wish to problematise their relationship from the perspective of nineteenth century studies – are these forms of media best treated as parts of the same cultural industry, sharing contributors, editors, and illustrators as they did; and if so, why are newspapers so often presented separately from journals and magazines? It is arguable that the press is the product of a single cultural industry in the nineteenth century, and is better studied as a whole.

This retrospective will next inform some constructive criticism of ongoing projects. Language barriers are an issue for searching and accessing records from sites, such as Europeana, that feature items in multiple languages. The problem of multiple languages is enhanced for media historians who need to search or browse in long runs, involving thousands of pages. More work is needed on the cross-national nature of the press, facilitating comparative work, and asking teams to consider the integration of translation software for those welcome digital media projects like Oceanic Exchanges and NewsEye that address the problem of titles in multiple languages by drawing on global titles in their corpora.

Building on an interest in the press itself, ongoing projects could also make it possible to interrogate corpora of data for characteristics of the media itself. For example, how did titles and publishing patterns change by year, by country, after 1850? How did issue prices or internal structure change over time for different categories of the press? How do the names and pseudonyms of correspondents and authors overlap with other titles? Furthermore, can these projects enable the creation of a single digital hub that allows researchers to see what has been digitised and what remains undigitised.

It would also be important for software for analysis of these corpora to include effective visualisations that users can themselves deploy to circulate their findings meaningfully. These will help users understand algorithmic input to and results from analysis, often reported statistically in graphs that are not always readable to humanist scholars. Both Impresso and Living with Machines seem conscious of users, and the desirability of including analysis and reflection in their projects. In general, consultations between projects and potential readers should happen much earlier in the conceptual and project building cycles, at a point where their input can be formative to project design and aims. Finally, sustainability should be foregrounded while at the moment it appears as a secondary concern.

6 Towards a Critical Framework for Digital Newspaper Scholarship

Paper author*: Andrew Prescott.*
Author institute*: University of Glasgow.*

I feel privileged to have witnessed the use of newspapers for historical research in Britain evolve from grappling with microfilm reels through to the wealth of resources currently available online. The way in which digital projects have opened up newspapers as accessible historical sources for a wide variety of topics is remarkable. While many scholars (including me) use fairly traditional research methodologies in exploring online historical newspapers, the way younger scholars are using innovative digital methodologies to explore such topics as social networking, history of ideas, and epidemiology is very exciting. As one of the most extensive international pools of historical data created by the first wave of the digital revolution, newspapers are already demonstrating the transformative potential of digital scholarship.

My own use of newspapers in historical research has been focussed on investigations of associational culture in eighteenth- and nineteenth-century Britain, and the availability of digital newspapers has been fundamental to this research. However, I have become increasingly aware of a number of issues which we need to address. I suggest that these questions might form the basis for a framework of critical newspaper studies. Some of these themes are:

1. What is a newspaper? Many existing newspaper digitisation projects were set up in response to immediate curatorial pressures, such as the British Library's need to vacate its premises at Colindale. The scope of these newspaper collections reflects long-standing pragmatic curatorial decisions and these require further exploration. Why should some periodical publications be included in digital newspaper collections and others not? What are the implications of the way in which collections have been divided for the structure and character of our digital corpora?

2. National boundaries. While the web is international, the way in which digitisation of newspapers has proceeded has reinforced national boundaries. National libraries have prioritised digitising newspapers from their own country. The British Library has focussed on its British holdings, but has paid little attention to digitising foreign holdings, even though some of these are unavailable in their home countries. Trans-national use of newspapers appear to be limited.

3. Fragmentation. Digitisation has proceeded on an institutional basis, so that runs of early newspapers are split across different packages. The split of the British Library's newspaper holdings between two packages, one not generally available to university researchers, is unfortunate. By contrast, microfilm coverage of newspapers was more systematically planned. What are the effects of this fragmentation of digital coverage?

4. What are We Dealing With? We have little information about how digitised collections such as the Burney Newspaper collection were formed. There are evident gaps in the digitisation coverage of particular collections and the implications of this are not clear. Above all, we are unclear as to how different collections relate to each other and what collections remain undigitised. As scholars increasingly use newspapers for innovative forms of quantitative analysis, a strong critical understanding of the nature of the newspaper archive will be essential.

Bibliography

A. Bingham. 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2):225–231, 2010. ISSN 0955-2359, 1477-4674. URL http://tcbh.oxfordjournals.org/cgi/doi/10.1093/tcbh/hwq007.

Rose Holley. A success story-australian newspapers digitisation program. *Online Currents*, 23(6):283–295, 2009.

Edmund King. Digitisation of newspapers at the british library. *The Serials Librarian*, 49(1-2):165–181, 2005.

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(4):E457–E465, 2017.

Ian Milligan. Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997â 2010. *The Canadian Historical Review*, 94(4):540–569, November 2013. ISSN 1710-1093. URL https://muse.jhu.edu/article/527016.

Clemens Neudecker, Lotte Wilms, Wille Jaan Faber, and Theo van Veen. Large-scale refinement of digital historic newspapers with named entity recognition. In *Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*, 2014.

L. Putnam. The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review*, 121(2):377–402, 2016. https://doi.org/10.1093/ahr/121.2.377. URL http://ahr.oxfordjournals.org/content/121/2/377.abstract.