Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-02668-4 OPEN

Reassembling digital archives—strategies for counter-archiving

Tobias Blanke[1]✉

Archives have long been a key concern of academic debates about truth, memory, recording and power and are important sites for social sciences and humanities research. This has been the case for traditional archives, but these debates have accelerated with the digital transformation of archives. The proliferation of digital tools and the fast-growing increase in digital materials have created very large digitised and born-digital archives. This article investigates how new digital archives continue existing archival practices while at the same time discontinuing them. We present novel methodologies and tools for changing memory and power relations in digital archives through new ways of reassembling marginalised, non-canonical entities in digital archives. Reassembling digital archives can take advantage of the materiality and the algorithmic processuality of digital collections and reshape them to inscribe lost voices and previously ignored differences. Digital archives are not fixed and are changed with new research and political questions and are only identified through new questions. The article presents six distinct techniques and strategies to reassemble digital archives and renders these according to three different types of new digital archives. We consider both the extension of archives towards evidence that is otherwise thrown away as well as the provision of new intensive, non-discriminatory viewpoints on existing collections.

[1] Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands. ✉email: t.blanke@uva.nl

Introduction

> "Today one can conceive (or dream) of recording everything, everything or almost everything […]." (Derrida, 1996, p. 74).

> '[E]very act of admitting data into the archive is simultaneously an act of occluding other ways of being, other realities.' (Bowker, 2014, p. 1797).

Recording everything is the dream of infinite archiving, as philosopher Jacques Derrida intimates in the quote above. While this has seemed impossible for a long time, the digital transformation of documentation and recording has made it seem not just possible but real. The archival literature has largely embraced this transformation. Digital archives are seen to be 'democratising' (Gauld, 2017; Taylor and Gibson, 2017) and present great opportunities for archivists (Cox and Students, 2007; Borgman, Scharnhorst and Golshan, 2019; Colavizza et al., 2021; Rakowski, Polak and Kowalikova, 2021). Moreover, digitisation is 'generating extraordinary collections of visual and textual surrogates' in a new 'process of representation' (Conway, 2015, p. 52). Everyone who has worked with digital archives knows how different they are from paper-based record-keeping. In fact, archives have developed a distinct meaning in computing and generally describe something else than they do in the tradition of archival sciences.

Digital archives can link to many different sites and their content can be changed to ever new representations. Consequently, they have become a much more encompassing term. Born-digital archives have grown fast in size and importance. Many of the institutions that are traditionally archived within systematic collections, such as governments and companies, have begun to work entirely digitally. But traditional archiving is also transformed by the digitisation of collections. Over the past decades, there have been numerous large-scale digitization efforts to make folders on archival shelves into data by computationally indexing them (Blanke and Kristel, 2013; Mordell, 2019). Archives have become 'data things' in Geoffrey Bowker's formulation supra, as their content can be read and transformed by computational machines.

However, the digital transformation of archives is also not as radical as some imply. Digital archives are not as universal as they appear at first sight (Carbajal and Caswell, 2021). There is still selection though often in a less formal way and with the help of or entirely done by algorithms replacing human curators. Not everything can be and will be recorded. There are still things that are excluded or at least not considered to be important enough for the new digital archives. In this sense, digital archives are not so different from their predecessors. Kim (2022, p. 531) sees biases continued by digital archives and exaggerated by their 'excess abundance'. They reproduce and accelerate existing biases, where some voices are privileged. For digital archives, there is a higher risk of new reinforcement bias, as users only see abundant digital records, assume they are everything and neglect other ones. Digital records are produced so excessively by powerful organisations like governments that other sources are easily ignored.

We start from the question how digital archives continue current archival practices while also discontinuing them. They can be misperceived as either a radical break from what has been done before or a simple continuation. Neither fully applies. The postcolonial scholar Ann Laura Stoler has proposed to stop thinking of the emergence of new historical formations as either 'too smooth continuities' or 'too abrupt epochal breaks' (Stoler, 2016, p. 6). For her, colonialism keeps on working, because it can deliver reactivations of past trends as well as new dispersions. Regarding archives and in particular a 'critical approach to the colonial archives', Stoler sees a research trend to read archives 'against their grain' (Stoler, 2002, p. 99) to reassemble marginalised and non-canonical perspectives. But she also warns us not to ignore what it is that makes an archive. We should start by reading 'along the grain of archives' and look for what she calls 'regularities' in archives.

To understand continuities and discontinuities for digital archives, we set off from our experience with digital research, where we permanently produce and reproduce different data representations. In the language of data science, its processes begin with data pipelines transforming one data representation into another. This is generally done to prepare data for more advanced computational processing, but these representations are also a form of what we call in this article reassembling the archives into different shapes. This article focusses on many as-yet-neglected novel representations we have produced in diverse data-science projects to explore what reassembling digital archives towards non-dominant knowledge means conceptually as well as in practice.

The article starts with a review of how the archival and related interdisciplinary scholarship addresses reassembling archives, introducing the distinction between the extensification and intensification of archival content. While the article is focused on practices of reassembling, in this section we also touch upon the wider debates around memory and power that motivate them. The second section discusses three cases of extensifying digital archives based on the results of diverse digital projects. Their results are re-approached from a new perspective by reconsidering their steps of data representations and understanding how we can reassemble them. In the next section, three cases of intensifying archives are presented and discussed. We conclude with a categorisation of the strategies used and what might be next steps for research on reassembling digital archives.

Reassembling digital archives

Archives have been a key concern of academic debates about truth, memory, recording and power. This is perhaps not surprising given the definitions that professional organisations offer. According to the *Society of American Archivists*, archives are '1. [m]aterials created or received by a person, family, or organisation, public or private, (…) and preserved because of the *enduring value* contained in the information they contain or as evidence of the functions and responsibilities of their creator, (…).' (Pearce-Moses and Baty, 2005, p. 30, emphasis added). Materials in archives are not just evidence sorted following archival principles but are also supposed to have 'value' that 'endures' beyond what is currently seen as important. Archivists are 'professional[s] with expertise in the management of records of enduring value' (Society of American Archivists, 2022).

With 'enduring value', archives make claims to sustained historical-political significance and invite discussions of their extensions in time and across space. Referencing 'evidence', archives are at the heart of what historian Carlo Ginzburg sees as the 'evidential paradigm' of research and reasoning with 'slender clues [which] have been adopted (…) as indications of more general phenomena' (Ginzburg, 2013, p. 124). Archives hold these 'clues' for the future and show what can be generalised. Until very recently and before the digital turn, they have almost had a monopoly on these clues. Now the clues are as widespread as the digital archives that hold them.

Claiming lasting significance and organising knowledge of the past into credible evidence, archives have seen broad interest. They have an important role in researching the past and

understanding the present. Archives are to many parts of the social sciences and humanities what the lab is to the sciences (Manoff, 2004, p. 13). But they are not just locations of research anymore. Following Ann Laura Stoler and others, the first 'archival turn' (Ketelaar, 2017) considers archives not just as sources for research but as 'subjects' of research, according to archival scientist Eric Ketelaar. Different disciplines study the multiple ways archives inscribe values and create evidence. As archives emerge through the selection, classification and ordering of documents of 'enduring value', they are also at the heart of social controversies and reinterpretations of the past.

In the Introduction to an agenda-setting special issue on 'archives, records and power' in *Archival Science*, the guest editors describe how they attempt to move away from the idea that archives are 'neutral repositories of facts' and rather places of 'social power' (Schwartz and Cook, 2002, p. 1). Earlier, Derrida had stated the importance of a political perspective on archives: '[T]here is no political power without control of the archive, if not memory.' (Derrida, 1996, p. 4). Documents are kept in the archives for a variety of reasons and are not a simple objective record of the past. With the first archival turn, archives have left their domains of archival and information sciences to become places of power to be excavated and explored by historical and social research.

In the same special issue, Stoler refers to Michel Foucault and his definition of archives as 'system of discursivity' (Foucault, 1982) to understand how they are shaped by 'selective forgettings'. Thus, archives are more broadly understood as 'a strong metaphor for any corpus of selective forgettings and collections (…)' (Stoler, 2002, p. 94), which are the result of political intervention. '[T]here is no state (…) without its archives', political philosopher Achille Mbembe reminds us (Mbembe, 2002, p. 23). But Mbembe also believes that archives can harbour traces of dissent, struggle and resistance, even as they have been set up as a system of discursivity of what can be said and what not.

The belief that archives can be different has motivated a second archival turn, concerned with 'using the archive(s) as a methodological lens to analyse entities and processes' (Ketelaar, 2017, p. 238). Ben-David (2020, p. 249), for instance, proposes 'archival thinking as an analytical framework' for social-media collections. In particular, she introduces 'counter-archiving' or 'building archives of Facebook that are designed to counter the platform's protocols of access to knowledge, that allow anticipating possible invisible connections, (…).' (Ibid., 254). Where there is archiving there can be 'counter-archiving' or 'reassembling' archival content to trace different narratives and work against selective forgettings. Stoler (2002, p. 109) also asks us to reassemble and 'create new archives of our own' by adding new materials and by reorganising existing ones in order to question how 'privileged knowledge' is produced and to rework the small differences that decide what is evidence and what has enduring value.

'Reassembling' is most famously associated with the work of the philosopher Bruno Latour. Although the word features in the title of his book *Reassembling the Social*, the actual concept remains somewhat elusive. Reassembling according to Latour can be best understood as 'deploying the sheer complexity of associations they [social scientists] have encountered' (Latour, 2007, p. 16). Reassembling digital archives could then describe and deploy differently the multitude of digital traces and associations. For archives, we can read Latour's reassembling the social as a suggestion to read along their grain and to look for and understand traceable associations. If what is 'to be assembled is not first opened up, de-fragmented, and inspected, it cannot be reassembled again.' (Ibid). Traces are reassembled to realise what is currently not in the archive or what is in there but is not seen or heard. Reassembled archives focus on how to create alternative histories from existing archival regularities. They emphasise what else can be said from archival collections without neglecting how they have been set up in the first place.

The historian Eric Hobsbawm has called for a new 'grassroots history' using counter-archiving and reassembling. Whereas Stoler mainly focuses on existing archives, he notes that grassroots sources are new to archives and need to be simultaneously made visible: 'Most sources for grassroots history have only been recognised as sources because someone has asked a question (…).' (Hobsbawm, 1998, p. 66). Such historical sources and generally all 'dark and community-built archives' (Guldi and Armitage, 2014) become visible through new questions and problematizations. Milligan (2016) considers the fundamental shift of historical analysis through very large digital archives, motivated by grassroots history. Digital archives allow concentrating on voices that have never before been part of historical writings in 'a massive documentary record of the lives of everyday people.' (Milligan, 2016, p. 85).

Hobsbawm is also ahead of his time, because he suggests that technology is key to researching such grassroots collections. Combining clues from these digital archives must rely on new techniques. In anticipation of the modern language of text and data mining, Hobsbawm tells us that traditional archival research finds the new by 'picking up diamonds in a riverbed', while grassroots history is 'more like diamond-or gold-mining, which require heavy capital investment and high technology.' (Hobsbawm, 1998, p. 66). Since Hobsbawm, technologies to reassemble the archives have turned out to be not as 'high tech' and require much less heavy investment. In data science, we deal with these technologies almost daily. They have become widespread with languages like R and Python and are reproducible with Jupyter Notebooks (Colavizza et al., 2021).

Following Hobsbawn, to reassemble also implies to question the boundaries of archives. What makes them different from archives that have not yet been assembled or have never happened? Such non-archives are collections that have not been assigned 'enduring value' or are seen as 'non-evidential'. The Washington State Archives, offering its opinion on non-archival records, 'has determined that these types of records do not have long-term value for public research.' (Washington State Archives, 2021). They can be destroyed before they are even submitted to the archive and without notice, thus leaving no trace and evidence. The idea of 'throwing-away' without consequence is also repeated in guidelines of the Canadian State Archive (2022) or archives at the Stanford University (2022). 'Throwing-away' belongs as much to archival places and practices as does the idea of 'enduring value'. Digital transformations, however, have meant that non-archival content is online just like archival one, thus blurring the distinction between non-archives and archives.

Questioning the boundaries of (digital) archives further implies not just an extension in scale towards more and more data things that would have been otherwise thrown away. It also entails intensive archival reassembling, (re-)making associations so that we can (re-)inscribe lost voices, places and stories. Our conceptual distinction between extensive and intensive reassembling is inspired by philosopher Etienne Balibar's work on extensive and intensive political universalism. For him, extensive universality means that a universal right to participation in politics includes everyone in a community, while 'intensive universality' challenges exclusions through the 'common humanity' that 'excludes exclusion' (Balibar, 2004a, p. 312). In another formulation, he distinguishes between 'universality as "inclusion" or "integration" (which I have called elsewhere extensive universality) and universality as "nondiscrimination" (which I have termed intensive universality) (…)'. (Balibar, 2004b, p. 46). Whereas Balibar is concerned with political processes, his

redefinition of universality and attempts to make it work while understanding its limits speak to our question of reassembling digital archives in non-exclusionary and non-discriminatory ways. Extensive reassembling of digital archives is about the inclusion of non-archives through grassroots research. Intensive reassembling works towards nondiscrimination by making new associations rather than expanding the scope of an archive.

To intensify and extensify digital archives, we need to turn to their forms of mediation. Generally, digital archives are guided not by the profession of archivists but by 'algorithmically ruled processuality' (Ernst, 2013). Algorithms enable decisions on what is part of the enduring value in digital archives and how our understanding of it can evolve. Archivists select in non-digital archives what is record-worthy and consider the limited space available on their archives' shelves to match their requirements of evidence and enduring value. To manage the ever-extending contemporary digital archives, a computer must take on the role of a mediator and custodian of knowledge. Reading along the grain of digital archives then also implies reading along the algorithmic operations that make digital archives. Reading against the grain should change this algorithmic operation.

In this article, we attempt to change the direction of the algorithmic processing that holds digital archives together. We start by decoding the algorithmic processing that makes digital archives through what Latour has called 'traceable associations'. Without their regularities, we cannot proceed to reassemble against the grain of digital archives. Unlike traditional archives, digital archives are not fixed, and their relationships with non-archives have become fluid. We can make non-archives into research materials through Hobsbawm's grassroots interests and with the appropriate algorithmic processing. In the following two sections, we explore first extensive reassembling of archives through things that are otherwise thrown away and then intensive reassembling and enabling new viewpoints on existing archival content.

Extensifying digital archives

There are many online collections of documents that are best described as unsystematic attempts to collect without an archival lens of enduring value and evidence. Nevertheless, they are algorithmically accessible and can thus endure. They are not archives in a traditional sense but rather document repositories set up to serve several different justifications and applications. Following Hobsbawm's suggestion, they become archives through an interest of research. We call these 'incidental archives' in that they are the result of other social and cultural processes. Their online content is generally not created directly, but algorithmically made or pulled out of often invisible, inaccessible data sources, from what is also called the deep web. For instance, governments publish some of their documentation online and make them accessible to fulfil promises of transparency. To understand how we can counter-archive such online content, surface hidden stories and reassemble them, we need to start from the same algorithms and techniques that are employed to create them. To this end, we invert this process by 'web scraping' what we find online and transforming it back into the data it has been made from.

Web scraping or the automated collection of content from websites has long attracted interest by researchers. Marres, Weltevrede (2013) approach scraping as a technique for social research and praise its open-endedness. According to them, scraping is an 'on-going process' and follows a 'commitment to research-as-process'. For many scholars, web scraping makes it easier to access large amounts of data (Lazer et al., 2009; Dogucu and Çetinkaya-Rundel, 2021; Luscombe, Dick and Walby, 2022) and is seen as a step towards a radical transformation of research, because it provides new ways of accessing 'trails of data' (Li, Zhou and Cai, 2021). Web scraping, however, is more than just a different type of access. Scraping opens digital materials to multiple transformations and allows to create new archives. In almost all the data-science projects, we discuss here, scraping has been the starting point of reassembling efforts.

Web scraping deals with content that is generally in a form that must be parsed, reformatted and reshaped to fit into digital archives. Static scraping accesses the content of websites directly, while dynamic scraping allows interacting with the embedded programming in the websites but requires advanced resources and is increasingly hard. Websites have now a lot of content that is not easy to access and takes extensive work with tools like Beautiful Soup and Selenium (Ruchitaa, Nandhakumar and Vijayalakshmi, 2023). Their content is often designed to work against web scraping, and a range of legal issues exist (Nigam and Biswas, 2021). In the United States, there have been many court cases questioning whether copyright is infringed by scraping. In Europe, national data protection agencies have been concerned with the violation of personal data by web scrapers. In India, unauthorised access to computers is against the law.

Online incidental archives generally consist of document collections that are made accessible through diverse forms of search algorithms. They are online archives of loosely connected documents, held together by search algorithms. The UK's 'Immigration and Asylum Chamber: Decisions on Appeals to the Upper Tribunal' (Tribunal Decisions, 2023) is a good example for an online collection of potentially important but also difficult-to-access documents. Hidden behind a generic search interface, are detailed descriptions of how a national asylum system deals with its cases. There are few better records of the daily experiences of asylum seekers struggling to make their case against a myriad of legal and administrative problems. Here, we also find recordings of many aspects of asylum seekers' daily lives. Legal cases are known to be one of the few records we have of otherwise forgotten or ignored groups' quotidian existence. To describe the experiences of common people since the 17th century, Hitchcock and Shoemaker (2006) have worked on the online publication of the Old Bailey records, the Central Criminal Court of England and Wales. Legal recordings are, however, not made to give voice but to present evidence in a case. They need to be reassembled to make quotidian experiences visible and intelligible.

The Upper Tribunal Decisions (2023) lacks all features of an archive, as it is not organised beyond simple search functionalities. There is no context to the decisions. Any archival organisation has been replaced by the 'algorithmically ruled processuality' (Ernst) of generic search with few additional possibilities to specify these searches. To reassemble it, we only have the option to break it down completely and download its documents in a brute-force version of web-scraping. In this case, we have used 'URL-hacking' or the transformation of URLs to provide direct access to the data. First the total number of cases is determined, to then go through all possible identifiers by transforming URLs to see whether a document is available. If that is the case, the document is downloaded and parsed for further processing. The Upper Tribunal Decisions (2023) is an extreme case where no usable metadata is available. It is as unworkable as an archive as it is useful for Hobsbawm's grassroots research.

Technically similar but about organisations rather than individuals is TED (2022), an archive 'allowing free electronic access to [the European Union's] call for tenders' documents such as contractual documentation, technical specifications, annexes, questions and answers etc.' It is made for 'Contracting Authorities' to ease publishing calls and for 'Economic Operators' as a single point of access. It is complemented by TED Tenders

(2022), which publishes award notices at the end of a procurement process. Compared to the Upper Tribunal Decisions (2023), these two EU archives are also document repositories but provide much richer access options and emphasize computational standards. They are built around a sophisticated so-called e-repository tool that is made not only to publish online documents but to help 'Contracting Authorities' and 'Economic Operators' with their procurements by EU institutions (TED, 2022). This means the search interface of these archives is powerful and customised to the needs of particular stakeholders. However, the repository's standardised formats have changed over time and curated by whoever has created its documents, which makes them difficult to parse.

With its additional technical capacities, TED (2022) can be reassembled in virtual collections (Bryant et al., 2015). The expert search functionality together with the consistent representation of content in the URL can be employed to narrow the search results via keywords like 'drones' or 'border' and additional metadata like 'status' or 'start date', creating a new view or virtual collection of documents of interests. Regular expressions, built into the search interface, help define these virtual collections by overcoming inconsistencies of spelling or uncertainty of keywords. The virtual collections slicing the archive into an area of interest can then be downloaded using restful web services (Anderson and Blanke, 2015). The downloaded documents are PDFs, which must be transformed into plain text in a non-trivial process. In several projects, we have tried to read against their grain by critically examining border technology industries and policies in the EU (for instance, and (Kirkeng, 2021)). Procurement documents are often the only clues we have for otherwise secret and opaque government practices.

Little structure and PDFs are typical for incidental archives, which are only accessible with significant technical investment and expertise. These sites are in fact themselves subject to archiving by the Internet archive. Web Archive Tribunals (2022), for instance, is the 23 April 2022 snapshot of a UK asylum tribunal decision from 2002. The Internet Archive deals with changes in websites by taking snapshots of them. Its web scraping is meant to provide a reproduction of the original site and not a reassembling. Compared to the previous two discussed archives, the Internet Archive, however, has an added temporal real-time axis, which allows for new types of reassembling. Whereas time is an additional metadata item in the already discussed archives, here time targets the archive itself as a reflection of changes to the underlying content. The archive is tracked real-time. Following its temporal grain directly allows us to observe historical activities and sample them live.

For the GUARDINT project (2022), investigating intelligence oversight practices in Europe, we have scrutinised the work of several NGOs on surveillance and oversight issues. Based on previous research within the project, seven NGOs have been selected: Amnesty International, Article 19, Big Brother Watch, English Pen, Liberty, Open Rights Group and Privacy International. For each NGO, the most relevant web page at a particular time has been identified to capture their activities and campaigns. Often, this is the homepage, but when the homepage does not contain enough meaningful content, other pages are scraped such as 'news', 'campaigns', 'blog', etc. Compared to the first two reassembling examples, we therefore need to understand more about the content changes and the relevance to the research question on surveillance practices. Thus, while the first two examples have reworked archives largely through technical expertise, this archive has been reassembled in a collaborative human-machine effort during a workshop in July 2022.

The NGOs' website snapshots are available through the Internet Archive's Way Back Machine using the Link Ripper Tool (2022), which collects a set of URLs from the Internet Archive including the timestamp. The corresponding web pages can be downloaded and their texts, hyperlinks and images extracted. Link Ripper's results are highly redundant, as the Internet Archive snapshots have often not changed much. To address this, we have created an algorithm to remove redundant pages that are 90% identical. Otherwise, we would be in danger of following digital archives' reinforcement bias because the same message would be considered again and again. Moreover, the Internet Archive captures vastly different numbers of snapshots per website, depending on how popular the site is. This means that we might lose sight of the contributions of smaller, lesser-known NGOs. We counter this with fairer temporal sampling. Four pages are the maximum number of per month we hold in case there are at least four samples. Otherwise, we use the maximum of available pages. Figure 1 shows the overview of available pages and their temporal distributions.

This section has shown how to reassemble online archives by re-collecting them from the web and modifying their algorithmic mediation. We have introduced strategies like virtual collections or exploiting the temporality of Internet Archives. With these methodologies, we have extensified digital archival collections and made digital archives more inclusive following diverse grassroots questions. The next section pays attention to internal transformations and intensifying content in archives by targeting the structure in their documents to surface different stories and voices. The aim is to make digital archives less discriminatory.

Intensifying digital archives

In this section, we move along archival grain not in terms of algorithmic access as previously but in terms of document structures, working with the files directly rather than on the level of collections. As before, we are guided by a research question to transform incidental archives. For the Upper Tribunal Decisions (2023), we are interested in the role of social-media companies for gathering evidence on asylum. Figure 2 shows how often particular platforms are mentioned over the last years and how social media gains almost exponential importance in the asylum cases. By 2021, they are in about 15% of all cases. To understand their growing importance, we have first followed the structure of the cases and identify relevant parts in them, as they are generally about many other things than just social media. To this end, we identify all sentences in the collection that contain one of the following: Facebook, Google, Telegram, Twitter, Viber, WhatsApp or YouTube. We have kept these as well as five sentences before and after, while removing sentences that overlap. This can be called issued-based reassembling, leading to a new collection of 1100 relatively short subdocuments.

With this subdocument collection, we proceed to read along the grain of digital archives following entities in the documents like names, organisations or places. Entities can be highly effective for counter-archiving at a micro-level. The *Freedom of Information Archive* enables access to classified US-government sources in a 'database of over 3 million documents about diplomacy and foreign policy.' To achieve its objectives, it uses named-entity-extraction, which extracts entities from documents 'to generate and test their arguments about diplomacy at the micro-level' (Connelly et al., 2020, p. 778). Fan and Presner (2022) employ similar techniques to reassemble existing Holocaust survivor testimonies for a micro-history of resistances.

With the entities, we proceed to build a knowledge base, which stores all the information as structured data that can then be used in further analysis or to create visualisations and make inferences. Knowledge bases can be seen as graphs where the entities in the text are nodes and the edges are relationships (Chiusano, 2022).

Fig. 1 Temporal sampling strategy for web archives. We sample a maximum of four snapshots per web page if there are four or more. Otherwise, we keep the maximum of snapshots.

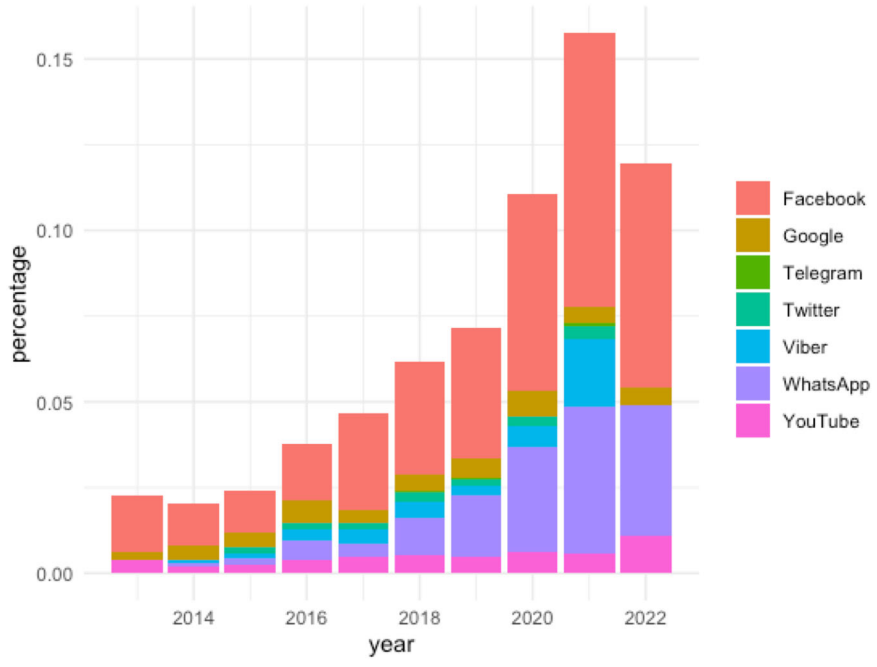

Fig. 2 Proportional count of social media platforms mentioned in Upper Tribunal Decisions (2023) subdocuments (until July 2022).

To create them, the computer must identify candidate entities for the nodes. In our case, this is a complicated, error-prone process, as the same entities are often described in manifold ways. The UK's 'home office', for instance, is also known as the 'Home Department' or 'UK home office'. We resolve all entities by checking whether they have a corresponding Wikipedia entry and keep the title of that entry as the identifier. In a second step, we extract the relationships between these entities using an end-to-end language model: REBEL (Huguet Cabot and Navigli, 2021) fine-tunes the BERT language model to achieve state-of-the-art performance in relationship extraction. We only keep relationships with entities about social-media companies.

Figure 3 shows a partial visualisation of the resulting network. The digital archive has completely been reassembled into a graph

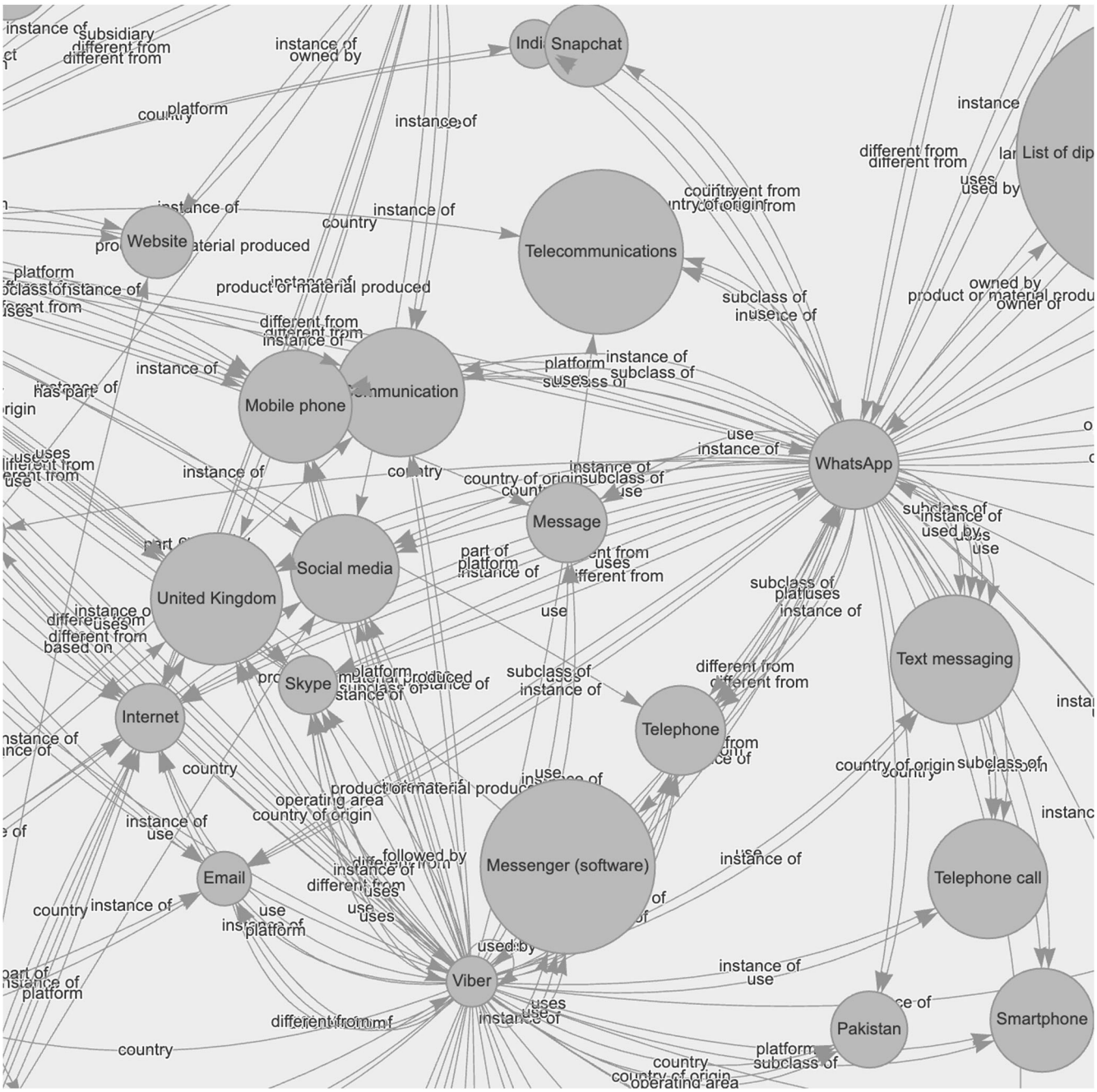

Fig. 3 Network visualisation of entities in Tribunal Decisions (2023) subdocument extracted using REBEL (Huguet Cabot and Navigli, 2021).

of social-media relations inviting us to retell the story of experiences in asylum applications through visual narratives. Unfortunately, this kind of advanced reassembling through new relations comes at great computational cost with the script running for several hours. The results should also not be read as the definite statement on the relations in the archives but as an incomplete but novel perspective on the archive. They show how language models can play a role in re-assembling archives.

Our second example of reassembling digital archives through their structure is based on an archive of power. The UK Parliamentary Archive contains the (digital) Hansard (Odell, 2021), verbatim transcripts of parliamentary debates. In the related data-science project, we have been interested in understanding how debates about security and intelligence agencies have changed over time and what kinds of alternatives have been articulated (Aradau, Blanke and Hussain, 2023). In particular, we wanted to find dissenting and alternative voices and topics in parliament that had been excluded. Here, we present data work that has not been included in the project's publications in order to showcase some of the many different data representations that are permanently (re-)produced in this typical project.

Given that security is a big topic in parliament, our enquiry has focused on the UK's Government Communications Headquarters (better known by its acronym GCHQ), which provides signals intelligence and information security to the government. In this case, we do not have to scrape the data ourselves but can reuse it from Odell (2021). His dataset includes every speech made in the House of Commons between the 1979 general election and the end of 2017. It employs among other things the pre-digital metadata and organisation of the Hansard records, which has

made it difficult to collect speeches only about the issue of 'GCHQ'. We decided to add an additional data cleaning step and kept only speeches which contained the keyword 'GCHQ' in the top 10% percentile of frequencies.

In the earlier example in this section, we focus on the entities and their relations using sub-documents. Here, we delve deeper into document structures and sentences and syntax as sources of reassembling the archive. Syntactic parsing is the automatic analysis of a natural language's syntactic structure, where sentence elements and their relationships can be discovered in a 'computational hermeneutics' (Mohr, Wagner-Pacifici and Breiger, 2015). We can determine which actions ('verbs') play a particularly important role in a parliamentary speech and which subjects and actors ('nouns') drive these actions.

Dependency parsing represents a sentence's grammatical structure and defines the relationships between so-called head-words and words, which modify heads. We attend to how actors act by zooming in on the combination of nouns and verbs in a sentence. A typical opening of a parliamentary speech such as 'May we have an early statement' can be represented through various new forms depending on the hidden relations of the text. They can be a list of the constituent words as well as lemmas like 'May – Pronoun – have …' or a multitude of other more detailed representations like the part-of-speech tags used in the Penn Treebank Project (Taylor, Marcus and Santorini, 2003): 'MD PRP VB DT JJ …' as well as many more. With these structures, we can ask specific questions that target relations between actors and acts via nouns and verbs. Figure 4 shows the most frequent noun-verb combinations in the parliamentary debates about GCHQ. Our reassembled archive contains legislative language, words that are embedded in arguments and lead to action. Points and opportunities are made, rights are claimed, and power is invoked.

Figure 4 shows what to expect in general in parliamentary archives if we follow frequency counts. To read against the grain of the Hansards and discover marginal subjects, we should investigate word relations that are not frequent using another linguistic relation. In the DOBJ-relation, '[t]he direct object of a verb phrase is the noun phrase which is the (accusative) object of the verb.' For instance, in 'The House votes for more money' 'votes' is a DOBJ-relation with 'money'. We have extracted all such direct object-object dependencies where the target has not been a very common word. This helps find new or different topics: In 1982 Cold War topics of classical 'espionage' are discussed, while in 1996 the idea to 'register paedophiles' appears. We have found that DOBJs give a very good overview of what makes an idea special in different historical situations. We can finally summarise this for dedicated concepts. Figure 5 shows a visualisation that focuses on DOBJ-relations around 'service' and 'right'. In this bubble graph, each bubble stands for one associated word and the size of the bubble for its frequency.

Both strategies of intensive reassembling operate on textual structures that we extract with advanced computational tools. The third method in this section targets 'paratexts' (Parrish, 2022), which in literary theory are texts surrounding the main body of

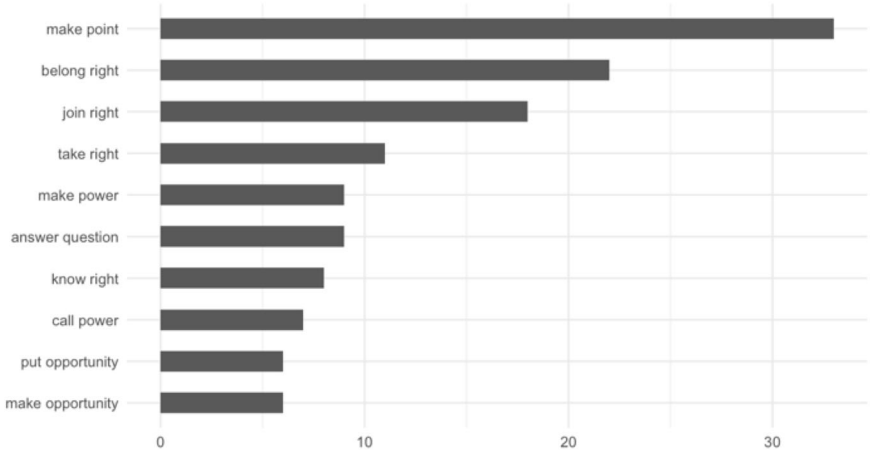

Fig. 4 Frequent noun-verb combinations in the UK parliamentary archives (Odell, 2021). The combinations are based on the automatic analysis of all sentences' syntactic structures'.

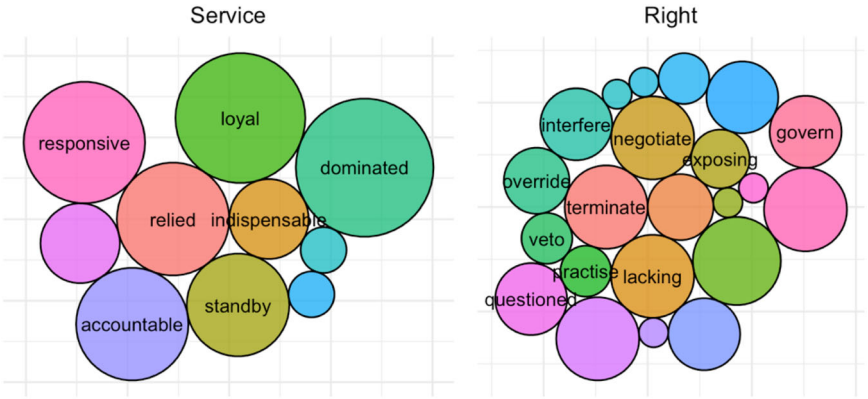

Fig. 5 Bubble graph of the DOBJ-relations in the Parliamentary Archives (Odell, 2021) for service and right, where the the direct object of a verb phrase is the noun phrase, which is the (accusative) object of the verb.

text. They frame the text as editor comments, dedications or back matters and relate texts to other texts that are from the same author or that are similar in other ways. Computationally, they can be automated summaries of the texts or in this case the keywords that shape the reception of the texts. They can be either used as summaries of the texts or as inputs into further computational processing. In particular, we are interested how we can use paratexts to summarise topical movements in a temporally indexed archive like the Internet Archive.

A very good way of summarising developments over time according to a set of documents are topic models, which relate words with each other without relying on existing structures. The topic modelling approach Latent Dirichlet Annotation (LDA) (Jelodar et al., 2019) assumes that each document belongs to several topics (k) with a certain probability, where each topic consists of several predefined keywords. We have used topic models in other projects to summarise historical developments and feed further data processing (e.g., Blanke and Wilson, 2017). However, interpreting topic models can quickly become like 'reading tea leaves' (Chang et al., 2009), as there is great flexibility in analysing them. In the earlier discussed project on NGO-websites from the Internet Archive, we have therefore developed a particular human-machine workflow that starts from the conceptual understandings of domain experts to make sure the topics are grounded.

Exploring the websites of NGO-campaigning against mass surveillance in the UK from the last section has been a human-machine collaborative effort, where interpretation by humans and machines alternate. It has started with selecting sites of interest within the NGOs' websites, but other steps of reassembling have also required humans and machines working together. For instance, we have decided to remove all words that do not appear in a list of 10,000 most common English words and eliminated spelling mistakes that are common in websites. The final collection consists then of about 20 m words. But the main collaboration has extended beyond data cleaning and into the analysis. Humans and machines have collaborated through 'seeded' or 'guided' topic modelling to avoid it becoming automated reading of tea leaves. With the conceptual guidance through humans, we could focus on alternative histories that might have been otherwise overlooked.

With seeded LDA (Jagarlamudi, Daumé and Udupa, 2012), topics are not only learned from the document collection but also pre-defined from a 'seed' of keywords. In regular LDA, first each word is randomly assigned to one of the pre-defined overall number of topics. Then, frequently co-occurring terms are collected together into a topic. After several iterations over all topics and words, the model converges. For seeded LDA, we can, e.g., seed the term 'data' to topic 1 by providing an extra bias for 'data'. While the model still converges by calculating the probabilities of words and topics per documents, the probability of 'data' belonging to topic 1 remains higher.

To create a human-machine workflow around seeded topic models, we ran a workshop with the GUARDINT project to decide on a set of topics and keywords to describe them by reading through a subset of the documents. In parallel, we ran several unseeded topic models to understand more about the overall distribution of keywords. The preselected topics corresponded to the qualitative research done in the project prior to the workshop. We iterated this interaction several times, comparing human-selected words with computed ones, before arriving at the following seven topics and their keywords:

1. *state_surveillance*: [state, mass surveillance, agenc*, home office, data, information, hack*, security, police, bulk data, interception, track*, database, spy*]
2. *corporate_surveillance*: [facebook, big tech, data, information, track, twitter, google, microsoft, amazon]
3. *general_democracy*: [democracy, rights, digital rights, civil liberties, freedom of expression, free speech, discrimination, chilling effect]
4. *singular_democracy*: [privacy, transparency, safeguard*, trust, data protection, rights, democracy]
5. *actors*: [public, press, media, journalist*, activist*, expert*, snowden, parliament, committee, commissioner, congress, government, wikileaks]
6. *resistance*: [whistleblow*, campaign*, scandal, petition, lobby*, report, media, press, court, leak*]
7. *oversight*: [scrutiny, oversight, snoop*, act, bill, tribunal, court, commissioner, committee, oversee, control]

With these seven topics, we can create more meaningful summaries of our time-indexed reassembled web archives that include human and machine understandings of its documents. Figure 6 shows two perspectives on the resulting timeline. At the top, we see how many times a topic has been the most important one in a year. This helps understand the main discussion points and how they change. Oversight became very important in the years after Edward Snowden had revealed the extent of global government surveillance, while questions of democracy and surveillance appear across the years. At the bottom, the Figure visualises how important a topic is for all sites in a year. It is more focussed on the distribution of the debates. Oversight, as we have defined it, peaks in 2016 and almost disappears in later years.

This concludes our discussion of reassembling through intensifying. The first project has tried to overcome the exclusion of refugee voices through new networked visual narratives. The second one has unlocked who the actors and what the actions are in parliamentary debates about GCHG. It has shown how to recover marginal contributions to the debates. In the third project, we could work against the discrimination of voices through topic models with a guided approach of human-machine interactions.

Conclusion: strategies of reassembling digital archives

The paper has presented several techniques and strategies to reassemble digital archives, taken from diverse data-science projects. Table 1 renders them following the three types of archives that we have discussed. It categorises them according to the grassroots research linked to the archives as well as how we can invert them. Extensification strategies for reassembling are shaded, while intensification ones are not shaded.

Legal cases can be reassembled into records of everyday experiences and the difficulties asylum seekers encountered making their claims for protection. While they are bound to attract strong grassroots research interests looking for underrepresented experiences, these are also examples of incidental archives that appear on the Web as mundane document collections. They generally do not have additional metadata and are made possible because generic search algorithms are cheap nowadays. To read against their archival grain, we have presented two strategies. The first one collects all materials and then rebuilds them from scratch into new archives. The second strategy employs issue-based search – focused in this case on social-media applications. Using issues, we can form subdocuments small enough to develop a visual narrative by extracting entities and their relations with a language model.

Also online are many examples of public transparency materials, our second example of incidental archival materials. These are often published by governments but increasingly also by companies and other organisations. Our research interest in understanding difficult to decode government policies on border

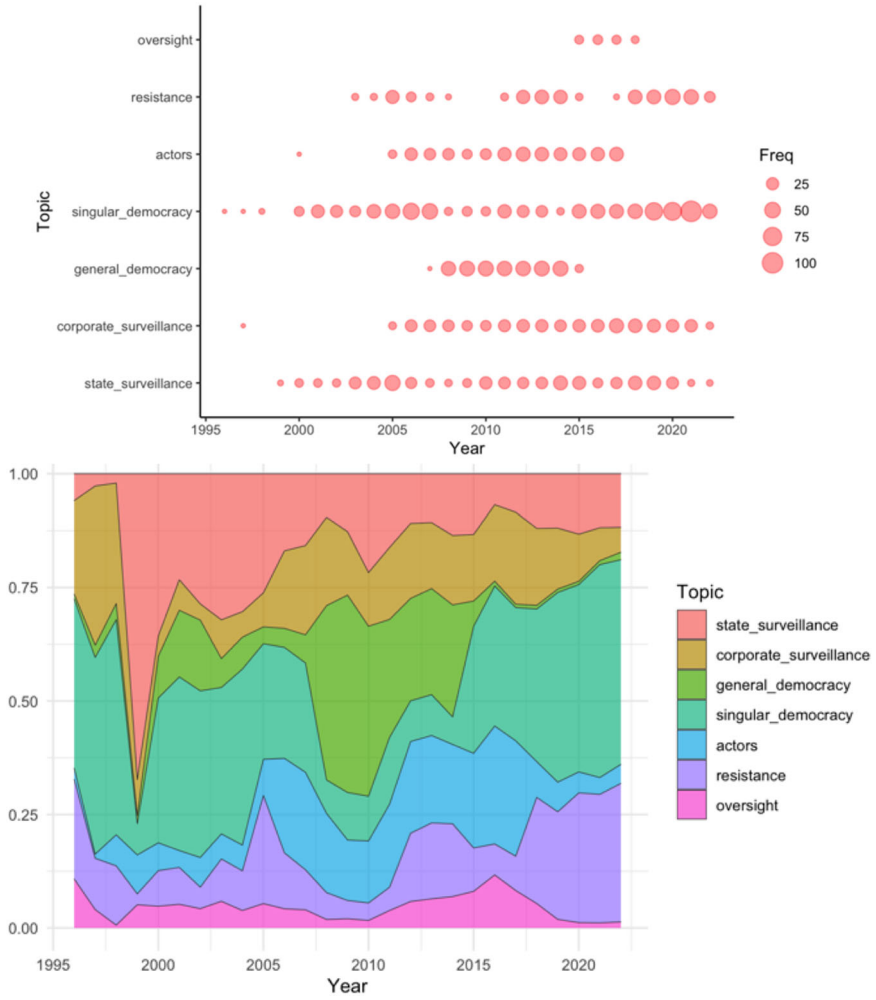

Fig. 6 Temporal visualisation of seeded topic models. The top shows how many times a topic has been the most important one in a year. The bottom visualises how important a topic is for all sites in a year.

Table 1 Strategies to reassemble digital archives.

Archive	Type	Grassroots Research	Along the Grain	Against the Grain
Legal Cases (Asylum)	Incidental Document Collection	Stories	Generic Search	Rebuilding the Whole Archive
	Subdocuments	Social Media Evidence	Issue-based Search	Visual Narratives from Entities and Relations
Public Transparency	Tender and Procurement	Access to Government Practices	e-Repository	Virtual Collection
	Verbatim Parliamentary Transcripts	Issue-based Democratic Discourse	Historical Metadata Scheme	Syntax Subjects and Actions
Web Archive	Real-time Archive	Monitoring Historical Activities	Seeding Sites	Temporal Sampling
	Time-indexed Archive	Unknown Histories	Seeding Keywords	Human-machine Interactions

technologies has led us to tender and procurement repositories. Compared to the legal documents, these are held together by a sophisticated dedicated search engine and digital repository software. To reassemble them, we form virtual collections that have allowed us to concentrate on the processing of specific subcollections. To read the grain of the second transparency archive, the UK's Hansard, we must follow its specific metadata developed well before the digital age. To research the development of discussions on GCHQ in them, we have broken them down into the most atomic syntactical subunit of sentences. This has allowed us to understand political relations or actors and actions. We have provided examples how to research what are typical actions for such an archive and how we could find historically and subject-specific actions.

Finally, we can read web archives with a temporal grain. They are built around snapshots, which we have used in two research cases for real-time archiving as well as historical indexing. The first case has allowed us to monitor and sample specific sites of NGOs to understand their political campaigning. To this end, we have developed a temporal sampling that provides a more even

distribution of larger and smaller NGOs than the Internet Archive. In the second example, we could exploit the temporal grain to index how the NGOs' discourses develop historically. We have seeded keywords to enable human-machine collaborations that describe these developments better than either machine or humans alone could have done. This has allowed us to derive paratextual topics summarising the historically changing focus points of debates on security within the UK parliament.

This article has argued that reassembling digital archives can work against dominant knowledge and attend to silenced or yet unknown voices. This means attending to the traces digital archives contain by reading not just against but also along their grain, as Stoler has suggested. By focusing on digital archives, we also challenged professional definitions of what an archive is. By including what Hobsbawn has called grassroots research interests, we could problematise what can be seen as an archive in the first place. In an increasingly digital world, there are many online collections which can become archives through new research interests and questions. To include both more archives and to work against the exclusion of voices in existing ones, we have recast Balibar's formulation of extensive and intensive universality. Table 1 summarises the strategies we have developed to extensify and intensify digital archives.

Table 1 also shows that reassembling digital archives remains partial and incomplete. There are many more archives to cover and many more non-archives to transform through grassroots research. Reading along the digital archival grain, we will be able to discover more ways to transform and extend digital collections or to reorganise their structures. There are more dynamic ways of online access to the archives rather than scarping and regular expressions, where we only access content as it is needed. The temporal reassembling of web archives currently relies on comparing the content of websites to avoid duplicates, but we could also include their hyperlinks. There are many more document structures beyond sentences and entities that would make it possible to explore digital document relations further. Exploring internal semantic relations within documents should be of particular interest. Other research strategies to make archival knowledge more inclusive through extensifying or to make it less discriminatory through intensifying could make better use of the often multi-modal nature of digital archives. For instance, in the case of the web archives, we have also collected images from the sites but have not used them in the end, because we have worked with topic models. Building on the archival turns, this article has offered a new perspective on data-science work as reassembling digital archives, which can look very different from traditional archives.

Data availability

The datasets of the Hansards and the EU document responsories for border technologies analysed during the current study are available from the repositories referenced through the cited project publications (Odell, 2021; Valdivia et al., 2022). The datasets of UK Tribunal on Asylum analysed during the current study are not publicly available due privacy restrictions but are available from the corresponding author on reasonable request.

Received: 19 August 2023; Accepted: 11 January 2024;
Published online: 02 February 2024

References

Anderson S, Blanke T (2015) Infrastructure as intermeditation–from archives to research infrastructures. J Doc. 71(6):1183–1202

Aradau C, Blanke T, Hussain I (2023) 'Making data visualizations, contesting security: digital humanities meet international relations', Global Stud Q 3(4). https://doi.org/10.1093/isagsq/ksad061

Balibar É (2004a) Is a philosophy of human civic rights possible? New reflections on equaliberty. South Atl Q 103(2–3):311–322

Balibar É (2004b) 'Racism, Sexism, Univeralism(s)'. In: N. Gorden (ed.) From the margins of globalization: critical perspectives on human rights. Lanham, Mayland: Lexington Books, pp. 43–61

Ben-David A (2020) Counter-archiving Facebook. Eur J Commun 35(3):249–264. https://doi.org/10.1177/0267323120922069

Blanke T Kristel C (2013) Integrating holocaust research. Int J Humanities Arts Comput 7(1–2):41–57

Blanke T, Wilson J (2017) 'Identifying epochs in text archives', in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2219–2224. https://doi.org/10.1109/BigData.2017.8258172

Borgman CL, Scharnhorst A, Golshan MS (2019) Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. J Assoc Inf Sci Technol 70(8):888–904. https://doi.org/10.1002/asi.24172

Bowker GC (2014) 'The theory/data thing: commentary', Int J Commun. 8(2043): 1795–1800

Bryant M et al. (2015) 'The EHRI project - virtual collections revisited'. In: L.M. Aiello, D. McFarland (eds) Social Informatics. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 294–303. https://doi.org/10.1007/978-3-319-15168-7_37

Canadian State Archive (2022) Non-archival record definition, Law Insider. Available at: https://www.lawinsider.com/dictionary/non-archival-record Accessed 18 Jul 2023

Carbajal IA, Caswell M (2021) Critical digital archives: a review from archival studies. Am Historical Rev 126(3):1102–1120. https://doi.org/10.1093/ahr/rhab359

Chang J et al. (2009) 'Reading tea leaves: how humans interpret topic models'. In Advances in Neural Information Processing Systems. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html Accessed 23 Jul 2023

Chiusano F (2022) 'Building a knowledge base from texts', *NLPlanet*, 24 May. Available at: https://medium.com/nlplanet/building-a-knowledge-base-from-texts-a-full-practical-example-8dbbffb912fa Accessed 15 Aug 2023

Colavizza G et al. (2021) Archives and AI: an overview of current debates and future perspectives. J Comput Cultural Herit 15(1):15. https://doi.org/10.1145/3479010

Connelly MJ et al. (2020) 'Diplomatic documents data for international relations: the Freedom of Information Archive Database', Conflict Manag Peace Sci p. 0738894220930326. https://doi.org/10.1177/0738894220930326

Conway P (2015) Digital transformations and the archival nature of surrogates. Archival Sci 15(1):51–69. https://doi.org/10.1007/s10502-014-9219-z

Cox RJ, Students TA (2007) 'Machines in the archives: technology and the coming transformation of archival reference', First Monday 12(11). https://doi.org/10.5210/fm.v12i11.2029

Derrida J (1996) Archive fever: A Freudian impression. Chicago, University of Chicago Press

Dogucu M, Çetinkaya-Rundel M (2021) Web scraping in the statistics and data science curriculum: challenges and opportunities. J Stat Data Sci Educ 29(sup1):S112–S122. https://doi.org/10.1080/10691898.2020.1787116

Ernst W (2013) *Digital memory and the archive*. Minneapolis, MN, University of Minnesota Press

Fan L, Presner T (2022) Algorithmic close reading: using semantic triplets to index and analyze agency in holocaust testimonies. Digital Humanities Q 16(3)

Foucault M (1982) The archaeology of knowledge: And the Discourse on Language. Pantheon Books, New York

Gauld C (2017) Democratising or privileging: the democratisation of knowledge and the role of the archivist. Archival Sci 17(3):227–245. https://doi.org/10.1007/s10502-015-9262-4

Ginzburg C (2013) Clues, myths, and the historical method. Johns Hopkins Press, Baltimore

GUARDINT (2022) Researching surveillance, intelligence & oversight. Available at: https://guardint.org/ Accessed 19 Jul 2023

Guldi J, Armitage D (2014) The history manifesto. Cambridge University Press, Cambridge

Hitchcock T, Shoemaker R (2006) Digitising history from below: the old bailey proceedings online, 1674–1834. Hist Compass 4(2):193–202. https://doi.org/10.1111/j.1478-0542.2006.00309.x

Hobsbawm EJ (1998) On history. New Press, New York

Huguet Cabot P-L, Navigli R (2021) 'REBEL: relation extraction by end-to-end language generation'. In Findings of the Association for Computational Linguistics: EMNLP 2021. Findings 2021, Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2370–2381. https://doi.org/10.18653/v1/2021.findings-emnlp.204

Jagarlamudi J, Daumé H, Udupa R (2012) Incorporating lexical priors into topic models. In *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. USA: Association for Computational Linguistics (EACL '12), pp. 204–213

Jelodar H et al. (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211. https://doi.org/10.1007/s11042-018-6894-4
Ketelaar E (2017) Archival turns and returns. In: A.J. Gilliland, S. McKemmish, A.J. Lau (eds) Studies of the Archive. Clayton: Monash University Publishing, pp. 228–268
Kim DS (2022) Taming abundance: doing digital archival research (as Political Scientists). Political Sci Politics 55(3):530–538. https://doi.org/10.1017/S104909652100192X
Kirkeng M (2021) Modelling datafication of borders using public procurement documents. Available at: https://dspace.uba.uva.nl/bitstreams/6e3841b1-60e6-453a-b589-8a1001264f20/download Accessed 21 Jul 2023
Latour B (2007) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford
Lazer D et al. (2009) Computational social science. Science 323(5915):721–723. https://doi.org/10.1126/science.1167742
Li F, Zhou Y, Cai T (2021) Trails of data: three cases for collecting web information for social science research. Soc Sci Comput Rev 39(5):922–942. https://doi.org/10.1177/0894439319886019
Link Ripper (2022) ToolLinkRipper. Available at: https://wiki.digitalmethods.net/Dmi/ToolLinkRipper Accessed 19 Jul 2023
Luscombe A, Dick K, Walby K (2022) Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Qual Quant 56(3):1023–1044. https://doi.org/10.1007/s11135-021-01164-0
Manoff M (2004) Theories of the archive from across the disciplines. Portal: Libraries Acad 4(1):9–25
Marres N, Weltevrede E (2013) Scraping the Social? J Cultural Econ 6(3):313–335. https://doi.org/10.1080/17530350.2013.772070
Mbembe A (2002) The power of the archive and its limits. In: C. Hamilton et al. (eds) Refiguring the Archive. Dordrecht: Springer Netherlands, pp. 19–27. https://doi.org/10.1007/978-94-010-0570-8_2
Milligan I (2016) Lost in the infinite archive: the promise and pitfalls of web archives. Int J Humanities Arts Comput 10(1):78–94. https://doi.org/10.3366/ijhac.2016.0161
Mohr JW, Wagner-Pacifici R, Breiger RL (2015) Toward a computational hermeneutics Big Data Soc 2(2):2053951715613809. https://doi.org/10.1177/2053951715613809
Mordell D (2019) Critical questions for archives as (Big) Data. Archivaria 87:140–161
Nigam H, Biswas P (2021) Web scraping: from tools to related legislation and implementation using python. In: J.S. Raj et al. (eds) Innovative Data Communication Technologies and Application. Singapore: Springer (Lecture Notes on Data Engineering and Communications Technologies), pp. 149–164. https://doi.org/10.1007/978-981-15-9651-3_13
Odell E (2021) Hansard speeches 1979–2020 Version 3.0.1, Evan Odell. Available at: https://evanodell.com/projects/datasets/hansard-data/ Accessed 19 Jul 2023
Parrish A (2022) Material paratexts, Allison Posts. Available at: https://posts.decontextualize.com/material-paratexts Accessed 19 Jul 2023
Pearce-Moses R, Baty LA (2005) A glossary of archival and records terminology. Society of American Archivists Chicago, IL, Chicago, IL
Rakowski R, Polak P, Kowalikova P (2021) Ethical aspects of the impact of AI: the status of humans in the era of artificial intelligence. Society 58(3):196–203. https://doi.org/10.1007/s12115-021-00586-8
Ruchitaa RN, Nandhakumar R, Vijayalakshmi M (2023) Web scraping tools and techniques: a brief survey. In 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT). pp. 1–4. https://doi.org/10.1109/ICITIIT57246.2023.10068666
Schwartz JM, Cook T (2002) Archives, records, and power: the making of modern memory. Archival Sci 2(1):1–19. https://doi.org/10.1007/BF02435628
Society of American Archivists (2022) Archivist. Available at: https://dictionary.archivists.org/entry/archivist.html Accessed 18 Jul 2023
Stanford University (2022) Archives and history office: What should You keep/what can you throw away? Available at: https://www.slac.stanford.edu/history/archnonarch.shtml Accessed 18 Jul 2023
Stoler AL (2002) Colonial archives and the arts of governance: on the content in the form. In: C. Hamilton et al. (eds) Refiguring the Archive. Dordrecht: Springer, pp. 83–102. https://doi.org/10.1007/978-94-010-0570-8_7
Stoler AL (2016) Duress: Imperial durabilities in our times. Duke University Press, Durham, NC
Taylor A, Marcus M, Santorini B (2003) The Penn treebank: an overview. In: A. Abeillé (ed.) Treebanks: Building and Using Parsed Corpora. Dordrecht: Springer Netherlands (Text, Speech and Language Technology), pp. 5–22. https://doi.org/10.1007/978-94-010-0201-1_1
Taylor J, Gibson LK (2017) Digitisation, digital interaction and social media: embedded barriers to democratic heritage. Int J Herit Stud 23(5):408–420. https://doi.org/10.1080/13527258.2016.1171245
TED (2022) eTendering. Available at: https://etendering.ted.europa.eu/general/page.html?name=home Accessed 19 Jul 2023
TED Tenders (2022) Contracts awarded by EU institutions - TED Tenders Electronic Daily. Available at: https://ted.europa.eu/TED/search/canReport.do Accessed 19 Jul 2023
Tribunal Decisions (2023) Immigration and asylum chamber: decisions on appeals to the upper tribunal. Available at: https://tribunalsdecisions.service.gov.uk/utiac Accessed 12 Mar 2023
Valdivia A et al. (2022) Neither opaque nor transparent: a transdisciplinary methodology to investigate datafication at the EU Borders. Big Data & Soc 9(2). https://doi.org/10.1177/20539517221124586
Washington State Archives (2021) What is a non-archival record? Available at: https://www.sos.wa.gov/_assets/archives/recordsmanagement/advice-sheet-what-is-a-non-archival-record-(march-2021).pdf
Web Archive Tribunals (2022) *Tribunal decisions*. Available at: https://web.archive.org/web/20220423214937/https:/tribunalsdecisions.service.gov.uk/utiac/2002-ukiat-4488 Accessed 19 Jul 2023

Competing interests

The author's work has been partly funded by AI4Media (Horizon2020, Grant agreement ID: 951911). The research for this article has been supported by collaborations on two projects: GUARDINT - Oversight and intelligence networks: Who guards the guardians? (ESRC, ES/S015132/1) and SECURITY FLOWS - Enacting border security in the digital age: political worlds of data forms, flows and frictions (H2020 European Research Council, No 819213).

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Tobias Blanke.