The Alan Turing Institute

Data Study Group Final Report: The National Archives, UK

9–13 December 2019

Discovering Topics and Trends in the UK Government Web Archive

https://doi.org/10.5281/zenodo.4981184

The copyright and database right in material produced by staff of The National Archives under this collaboration is Crown copyright or Crown database. Free and flexible re-use of Crown Copyright material is granted under the terms of the Open Government Licence (OGL), and Crown Copyright and database right material is published under the OGL.

Contents

This report has a companion GitHub repository containing example code: `https://github.com/alan-turing-institute/DSG-TNA-UKGWA`

1 Executive summary

1.1 Challenge overview

The National Archives is the official archive and publisher for the UK government and for England and Wales. It is the guardian of some of the country's most iconic documents and collections, dating back over 1,000 years to today, and including those published on the web by UK government departments and bodies.

The UK Government Web Archive (UKGWA)[1] is a vast resource of government websites and social media content, and an important source of recent national history, spanning 23 years. It contains over five billion resources [or distinct Uniform Resource Locators (URLs)] and is one of the most heavily used web archives in the world, serving hundreds of thousands of page views each month. The National Archives has a remit to preserve government-owned web content in all its forms (including web pages, official publications, datasets, social media, such as tweets and multimedia) and seeks to preserve this part of the record in its original context, wherever possible, through this archival resource. The UKGWA[2] is free to use and fully accessible via the web, and includes a full-text search service.[3] The size, variety of formats, and complexity of this vast collection makes the discoverability of its content challenging to the user and therefore puts great pressure on the search service to meet the needs of those users.

The Alan Turing Institute[4] is the UK's national institute for data science and artificial intelligence, with headquarters at the British Library. Data Study Groups are intensive five day collaborative hackathons hosted at the The Alan Turing Institute, which bring together organisations from industry, government, and the third sector, with multi-disciplinary researchers from academia. The National Archives was the Data Study Group Challenge Owner; their experts were present during the week, and are co-authors of this report. They provided the real-world challenge to be tackled by this

[1] https://www.nationalarchives.gov.uk/webarchive/

[2] https://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf

[3] https://webarchive.nationalarchives.gov.uk/search/

[4] https://www.turing.ac.uk/

group of researchers led by the Principal Investigator and Facilitator. This report is the culmination of that process.

The challenge we address in this report is to make steps towards improving search and discovery of resources within this vast archive for future archive users, and how the UKGWA collection could begin to be unlocked for research and experimentation by approaching it as data (i.e. as a dataset at scale). The UKGWA has begun to examine independently the usefulness of modelling the hyperlinked structure of its collection for advanced corpus exploration; the aim of this collaboration is to test algorithms capable of searching for documents via the topics that they cover (e.g. 'climate change'), envisioning a future convergence of these two research frameworks. This is a diachronic corpus that is ideal for studying the emergence of topics and how they feature through government websites over time, and it will indicate engagement priorities and how these change over time.

The National Archives last embarked on a project to use natural language processing (NLP) tools on the UKGWA in 2010.[5] It produced promising results and highlighted the approach as being useful in addressing some resource discovery challenges in searching unstructured content gathered over a long period of time. Ultimately that project ran into difficulties at the querying stage due to scale and did not deliver a user interface suitable for researchers. However, much was learned by the team involved and a great deal of progress has been made in the area in the intervening period in terms of software, computational methods, and increased availability of the necessary compute power to deliver the service to users. It was therefore considered timely to revisit the concepts with the ultimate aim of delivering improved access to the archive.

1.2 Data overview

The data consists of plain-text collections derived from HyperText Markup Language (HTML) pages. The HTML pages, drawn from the UKGWA, cover a very broad range of topics, from health to international policy, as the government influences all aspects of society. The datasets were

[5] `https://webarchive.nationalarchives.gov.uk/20101011131758/http://www.nationalarchives.gov.uk/documents/research-enews-june-2010.pdf`, page 4

created by selecting annual snapshots taken of these websites on the date closest to 1 January each year from 2006 onwards. In certain cases this covers the entire lifespan of sites. These snapshots were created through a combination of shallow and deep web crawling, and extracting the plain-text from the resulting HTML. Each page of each site is stored in an individual text file, organised by site and date of snapshot.

1.3 Main objectives

The overall objective is to use these curated datasets of reference documents to examine algorithms that are capable of identifying similar documents across the corpus and of inferring the topics they cover. This work will contribute to generating an overview of the contents of the UKGWA, which will be developed for inclusion in user-facing services, enhancing search and further enabling the use of the UKGWA.

The main aims are to give insight into 1) what approaches can be used to assist the understanding of the data within UKGWA 2) what are the most viable approaches to improving the resource discovery services offered to users.

1.4 Approach

The organisation of the UK government web estate has changed over time, which has a bearing on both the structure and the content of the archival data. Issues that are particularly challenging in web archive research (e.g. duplication, and changes in the content and technical make-up of the source websites over time) add complexity to the analysis and require contextualising information for researchers who are new to applying data science methods to web archives.

In order to improve the search functionality of the archive, we leveraged three central pillars of current research in information retrieval (see for instance Singhal [2012], Mitra and Craswell [2017]):

- Enriching the corpus with annotations to entities with unambiguous meanings

- Designing a preliminary interface for comparing the relevance of entities and concepts over time
- Providing a preliminary semantic search functionality to the user

To identify all entities mentioned in the corpus we employed spaCy, a named entity recogniser.[6] We then grouped mentions of the same entity (e.g. 'Obama', 'Barack Obama', and 'President Obama') using a database of aggregated statistics derived from Wikipedia.[7] We then built a prototype interface which would allow users to search the UKGWA both for entities and for single concepts, and to compare their frequency over time.[8] Once the user decides to focus on a specific concept or entity in a specific year, the search interface offers three ways of navigating the retrieved documents:

- Semantic search: the most relevant documents (as measured by Doc2Vec Le and Mikolov [2014]) are sorted based on their semantic similarity to the user query
- Cluster-based search: documents are grouped in semantic clusters (by applying hierarchical clustering) which may aid the user in meaningful browsing of the collection
- Entity-based relational search: the most closely related entities (as provided by the spaCy named entity recogniser) to the user query are provided, filtering the returned search results

1.5 Main conclusions

This series of experiments demonstrated how discovery and search of web archives can be usefully enhanced from what is widely available currently. However, both the discoverability and comprehension of search results may be improved still further by using a combination of different

[6] `https://spacy.io/`

[7] `https://github.com/fedenanni/Reimplementing-TagMe` and also the scripts `spacy_preprocess.py`, `tagme_preprocess.py` and `multithread_calc_entities_spacy.py` in the companion repository `https://github.com/alan-turing-institute/DSG-TNA-UKGWA`

[8] See the script `demo_interface.py` in the companion repository `https://github.com/alan-turing-institute/DSG-TNA-UKGWA`

techniques through a novel navigation interface. Our user-centred approach was informed by the experience of The National Archives' experts, in turn informed through user surveys (while The National Archives do not publish the comments received, some summary data is available).[9] They focused the team's efforts on the evaluation of the potential approaches for improving search and discoverability, and the promotion of transparency, as demanded by trusted institutions. The achievements of the week, and the areas identified for future work, demonstrate a notable change in machine learning capability since The National Archives last conducted experiments in 2010.

The team's work over the week showed that enriching the collection with named entities is possible and advances the browsing and discovery experience. There was additional promise in adopting topic based search (semantic search): the Data Study Group showed that it may achieve more accurate results and thereby further improve the user experience. Overall, we showed how a variety of NLP approaches could support The National Archives in offering more targeted access to the collection. Beyond the week's experimentation with sampled data, the aim would be to employ the suite of tools, techniques and approaches on the full scale of the UKGWA dataset. Equally, these methods are applicable to other web archives, or large textual corpora.

1.6 Limitations

The focus of this Data Study Group was on text-based content. While this is the most significant content in the UKGWA, non-text resources (images, videos and audio) are an increasingly significant part of the government web estate and therefore the UKGWA. Our work noted the negative impact of web pages reusing common content (e.g. navigation) on the results, but it was outside the scope of the week to remedy this.

Considering topic modelling, the work presented here is simplified, and does not account for the semantic change of words over time. Today, a 'tweet' is likely to be a piece of social media as much as it is a bird noise. A further limitation is that infrequent, niche or technical terminology

[9] https://www.nationalarchives.gov.uk/about/our-role/transparency/our-public-services/

may be suppressed by the larger corpus of more general text, hampering discoverability of these documents. The opposite problem also exists, that if a user searches using a term that does not occur in the UKGWA data, it is ignored, even if the topic is discussed using different words.

1.7 Recommendations

Given further research and development, there is potential for The National Archives to make enhancements to the existing search and discovery capabilities for the UKGWA. This may be achieved by further development of the existing services, developing bespoke services outside of UKGWA but using existing APIs, or creating standalone services based on extracts of its data. The National Archives team have a wide variety of data that can be used to impart context and structure on the data. Therefore, having the potential to give the user more control and power in navigating the archive's content, and to understand its structure. Understanding the needs of user communities is critical, not only existing users looking for specific content but also researchers who wish to explore the UKGWA as data.

Further research is needed to improve and extend the data science methods for enhanced impact and performance. The application of current state-of-the-art pre-trained deep language models [Devlin et al., 2018] could enhance semantic information retrieval and entity linking. Additional information sources, for example the creation of a domain knowledge graph starting from topical information from Wikipedia (based on how articles link to each other), would add more context and improve the accuracy of search results. This, combined with specific domain data from The National Archives, the application of existing data in structured formats, and making other knowledge sources machine-readable, would offer a transformative service to UKGWA's users. This could be further enhanced by the integration and extensions of systems for fast parallel research over large-scale collections of vectorised semantic documents (see for instance Johnson et al. [2017]).

2 Quantitative problem formulation

Ahead of the Data Study Group week itself, a set of seven research questions were developed to stimulate potential research directions to the research group. Based on the given time frame and skill set of the group, some of the questions have been addressed more thoroughly (1-3) over the week than others (4, 6, 7), and two questions will be used to shape ideas for future research (4, 5). They were:

1. How can we enrich each document with semantic information relying upon out-of-the-box Natural Language Processing (NLP) tools?
2. What unsupervised machine learning can be used to aggregate documents in thematically similar clusters?
3. How can the resulting metadata be used to inform description of the nature of the information they contain and guide the interpretation of categories?
4. What methods can track the emergence and evolution of topics across time?
5. What approaches can be used to differentiate between the functions and aims of government departments as expressed in individual websites?
6. How do we best explain the data science methods used on the UKGWA to its readers and users?
7. Can we develop workflows to aid the interpretation of any machine learning algorithms we develop, and encourage engagement with their strengths and limitations?

3 Qualitative problem formulation

3.1 User types

The purpose of this challenge is to explore algorithms that have the potential to enable further and purposeful use of UKGWA by end users. To that end, we dedicated time to tentatively scope and describe potential

personas and needs of end users, in order to anchor the variety of algorithmic approaches we explore in concrete use cases. The UKGWA is a very large collection that is actively curated by a team of experts, while being freely and fully accessible to its users, making it an extremely valuable resource. However, the size and its complexity make it difficult for researchers to use to its full potential. Although it is quite different and separate from other collections belonging to The National Archives, existing user research into users of The National Archives' online catalogue Discovery[10] brings useful insights on user behaviour.

Evidence collected through (unpublished) interviews and surveys during 2015 indicate that a 'steep learning curve' exists for new users; however, once the user has familiarised theirself with the system, they can build a mental model of it and can use the service successfully in their research efforts.

When we come to large-scale computational analyses of the web there are different levels of desired interaction, depending on the requirements and interests of the user. Some will just want to see results, maybe a visualisation of frequent words and phrases, or a network diagram. Others will want to understand the provenance of the results they are seeing, which might include algorithmic explanation. This is a particular challenge with the advent of data processing algorithms which are more complex and less generally understandable and introduce an element of uncertainty and chance into the results.

3.2 Algorithmic transparency and trust

Without knowledge of the capture process, an archived web page looks like any other web page. Transparency and explanation is needed not only to improve user experience, but also to enhance confidence and trust in the web archive.[11]

Therefore, an important obligation is to provide an explanation of the processing and outputs that are maximally understandable to the public, and communicate the complexity and uncertainty in the data processing

[10] http://discovery.nationalarchives.gov.uk/

[11] https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf

and results. However, in here lies a delicate balance. Alongside the need for transparency and openness regarding its limitations, care must also be taken not to overstate the challenging aspects inherent in the collection, which could deter use of the archive or undermine the sense of skill or mastery a user can develop given some time to explore it.

Transparency, nonetheless, is a critical prerequisite for users to develop trust in the archive's search, exploration and presentation of results. It is also a recognised issue amongst humanities scholars that researchers often do not understand the provenance of the results they receive from search engines [Kemman et al., 2013]. Currently, search results from the UKGWA are not ordered by relevance; if they were in future, the interface would need to explain that ordering and how 'relevance' had been determined. In some situations this could influence the choice of algorithm but there must be a sensible balance between transparency and functionality. From the data science perspective, The Alan Turing Institute has published guidance on algorithmic transparency and trust, in the form of a guide for the responsible design and implementation of AI systems.[12]

4 User experience

4.1 User journey

For the purpose of this report, we assume that UKGWA users bear some resemblance to the Discovery users, and on that basis, suggest some potential user personas. Therefore, we have assumed the UKGWA user needs fall into two broad categories, which we term **advanced search** and **quick search**. Indeed, this is supported by user research conducted specifically on UKGWA, which led to the formation of personas such as 'professional services', which may be an example of an advanced search user, and 'member of the public' as a potential example of the latter. Any developments would need to support the needs of these personas.

Quick search will follow some or all of the following journey:

[12] https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety

1. Enter a **search term** which consists of a single word, or multiple words
2. Receive a list of **ordered** results (for some definition of ordered)
3. If the desired page has been found stop, or
4. Use the returned results to inform/refine a further search

Advanced search will follow some or all of the following journey:

1. Start with a **search term** which consists of a single word, or multiple words
2. Receive a list of **ordered** results (for some definition of ordered)
3. The returned list is annotated with key **entities** which were previously identified and extracted from the archive content
4. Select one or more of these **entities** to:
   (a) See a visual representation of how the predominance of those entities changes over time
   (b) Explore an interactive visual graph of related entities, where similarity is driven by co-occurrence in the archive. These discovered entities can then be used to start new searches
   (c) Explore an interactive visualisation of documents most similar to document that best matches the current entities

4.2 User interface

Figure 1 illustrates the results of an initial query issued by both **quick search** and **advanced search** users.

A linear list of results are presented, ordered by **relevance** to the search terms. Each result consists of the document title, augmented with context information, including the originating **department** and **date**.

Each search result is also assigned **semantic tags** which we here call **entities**. These are key terms which have been filtered by a match against Wikipedia (or another third party knowledge source) as a proxy for validity as a concept, person, location or object.

Figure 1: Wire-frame mock-up showing results of an initial query

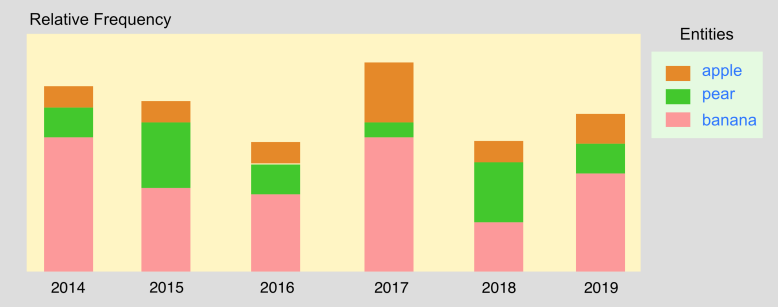

Figure 2: Wire-frame mock-up showing the relative frequency of entities in search results

These **entities** can be clicked to navigate to the following exploratory visualisations.

The wire-frame in figure 2 represents a chart presenting the prominence of terms over time. These terms are those which were selected from the search results prior to opening this visualisation. This meets a key user need to understand how focus on topics has evolved over time. **Advanced search** users will require further explanation of what is meant by 'Relative Frequency', and this is covered later in this report.

The mock-up in figure 3 represents a network graph of related **entities**. The links between **entities** are strong, bringing nodes closer together, when the entities at each end are closely related in terms of **similarity**. How similarity is defined will depend on the algorithms used and the types of entities. For example, co-occurrence of entities could be a sufficient similarity metric if those entities are words or people, while

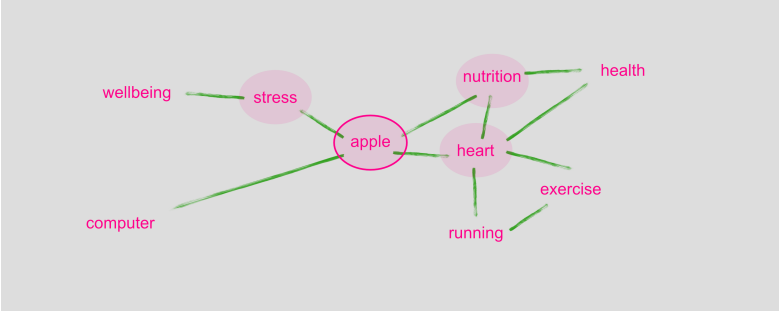

Figure 3: Wire-frame mock-up showing the graph of related entities

cosine similarity would be more appropriate for vectorised representations of words and documents.

These graphs will often be very busy, so the ability to filter by number of nodes is essential. This filter could be based on entity frequency, or the nearest neighbours to an entity of interest. The chart is interactive, in that it allows users to move the nodes around so that clusters can be viewed more clearly, and users can also double-click a node in order for it to be centred on the chart. After each user-initiated movement, the graph is adjusted using a mechanical model of attraction and repulsion. The popular D3.js[13] force-directed graph[14] is a good basis for such a visualisation.

Instead of a network representation, we can also visualise documents in a 2-dimensional vector space, as shown in figure 4. This depicts a view of documents that are related to the one that best matches the user's query, and specifically, its semantic tags. This view can be considered as a cloud of points. Clustering algorithms can be applied to label and to separate groups of documents in this vector space. This labelling is represented by the varying colours of the nodes in the picture.

4.3 User-facing explanations

For users who require explanation, they will need to understand how entities are represented in the space and how they are linked together,

[13] `https://d3js.org/`

[14] `https://observablehq.com/@d3/force-directed-graph`

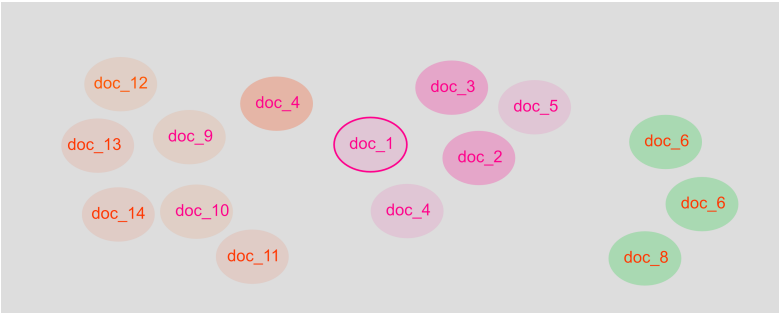

Figure 4: Wire-frame mock-up showing the documents in vector space

i.e. what is the similarity metric and how is it calculated? The following sample passages are suggested as explanations of the data processing and results to aid user understanding.

Quick search results, explanation:

> The search **query** text, which you have typed into the search box, is used to find documents that match. Those matching documents are shown on the results page.
>
> Those that match best are higher up the search results list. This works in a very similar way to the internet search engines which you might be familiar with.

A more detailed explanation for **advanced search**:

> Documents are indexed by separating their text content into word tokens. This would result in a very large dictionary of tokens of varying usefulness. To remove less useful tokens, they are filtered by matching against Wikipedia. If a Wikipedia entry exists, that token in considered to be valid and retained, and referred to here as an **entity**.
>
> When performing a search using query text, matching only occurs using the entities which appeared on Wikipedia. This means that if a search word does not have an entry on Wikipedia, it will not be used as part of the search.
>
> Search results are ordered by how well the documents match the **entities**.

An explanation of the **trends over time chart** for **quick search**:

> The chart shows how frequently topics occur over time. More than one topic can be compared, and this results in a stacked bar chart.
>
> Note that if a document appears over a period of years, even if unchanged, the matching **entities** will still be counted for each year.

A more detailed explanation for **advanced search**:

> The chart shows how frequently topics occur over time. More than one topic can be compared, and this results in a stacked bar chart.
>
> The height of the bars is the frequency, per document, with which the entity appears in the archive for that year. A matching document will only contribute a count of one, and so multiple entity occurrences do not boost the match. A height of 0.05 means that a term matches 5% of the documents in that year.

The **network graph of linked entities** can be explained to **quick search** users as follows:

> The chart shows topics that are related to each other. The closer the topics on the chart, the more related they are in the archive.
>
> You can follow links from one topic to another to explore related topics and discover topics that you might not have known were related.

A more detailed explanation for **advanced search**:

> The graph of links and nodes shows how entities are related to others. The nodes represent the **entities**, previously matched against Wikipedia entries.
>
> The graph is automatically arranged by modelling how strong links are between nodes. This strength is the similarity between entities, which here is determined by the distance

between them in a vector space created using the widely used **Doc2Vec** algorithm run against the documents.

Because the **Doc2Vec** algorithm is non-deterministic, that is, the algorithm involves some randomness, the resulting output can vary even with the same data as input. You should take the resulting graphs as indicative.

5 Data overview

5.1 Dataset description

The dataset consisted of 3,893,092 files, totalling 23.9 gigabytes (GB). While it would be usual to perform exploratory data analysis prior to performing machine learning, it was decided that a better use of time would be to run an expert-led session describing the data. The dataset was a derivative extract which had been defined by The National Archives and was therefore already understood to a deeper level than summary statistics would provide.

In this session, The National Archives' experts described the processes behind populating the archive, how the datasets were created and where the challenges lie for both the archive and its users. This session then led to a wider discussion of the problem space and potential research avenues, and it was felt that this was particularly beneficial for the team as they started their experimentation with a stronger understanding of what we wanted to achieve.

The final resource presented at the event consisted of four datasets. Each is a plain-text collection derived from HTML pages in the UKGWA:

- **Dataset1** Government Hub Websites 2006–2019: pages from direct.gov.uk and www.gov.uk, throughout this period, with JavaScript removed by machine learning
- **Dataset1-raw** Government Hub Websites 2006–2019: pages from direct.gov.uk and www.gov.uk, throughout this period, without any pre-processing

- **Dataset2** A subpart of Dataset1. Thematically Sampled Websites 2006–2019: pages from approximately 450 government websites, which were pre-selected through high-level topic modelling [using MAchine Learning for LanguagE Toolkit (MALLET)][15]
- **Supplementary** Home Pages 1996-2019 and Blogs 2013-2019: every home page captured since the beginning of the UKGWA, which was the dataset used to seed the creation of Dataset2. Shallow crawls of government blogs, although most blogs are not crawled every month

The websites from which these datasets were drawn address all areas of life and activity influenced by UK government work. The subset of websites in Dataset2 were selected to reflect a range of these activities identified through prior topic modelling as relating to the Olympics, healthy eating, regional development, climate change, and statistics and transparency.

Dataset1 and Dataset2 were created by selecting snapshots taken of specific websites on the date closest to 1 January of each year, from 2006 onwards. Websites identified were then crawled to sufficient depth to capture most of their content, with sitemaps being used to identify pages in the domain when available. In many cases this range of snapshots covers the entire lifespan of sites. The home pages dataset was included partly because it was used to select Dataset2, and also because it could be used for a fine-grained analysis of topics over time. The blogs provided larger amounts of text material, but also a different type of content which was more likely to be driven by topical events from the outside world. For each dataset the text was extracted from the HTML using the Python[16] library Beautiful Soup[17], and individual text file (named according to the URL), and organised by website and snapshot.

5.2 Data quality issues

The pre-processed data files contain the body, header and footer of the corresponding pages, along with residual text resulting from the extracting

[15] `http://mallet.cs.umass.edu/`
[16] `https://www.python.org/`
[17] `https://www.crummy.com/software/BeautifulSoup/`

process. The headings in the file contain information about the structure of the websites; this content can be used in the future but is not relevant for computing the entities at this moment. However, files are stored in plain-text format, and to seperate the page body from header and footer content is challenging, as the HTML tags making this explicit were removed during the extraction. A more natural data format for web archive data is the Web ARChive (WARC)[18] file format, which combines multiple digital resources into an aggregate archival file together with related information. It provides better support for harvesting, access, exchange of data and has been the predominant format for web archives to present. However, it was decided for this Data Study Group that the volume of data in WARC format would either be too large to provide a reasonable subset of the archive, or would require significant pre-processing during the week.

The work was focussed on documents containing larger amounts of text, such as blog posts, news articles, speech or similar content. These would be more likely to be entity rich and therefore be more suited to document similarity techniques than home pages which are light on entities and heavy on navigation.

6 Experiment: entity recognition and disambiguation at scale

The first objective of the Data Study Group was to semantically enrich the collection, by automatically attributing entities and concepts to words with unambiguous meanings. This was done following current entity-centric approaches in information retrieval systems (see for instance Dalton et al. [2014], Nanni et al. [2017]), which embody the 'Things not Strings' vision.[19] Such a step would allow for a better understanding of the collection. Moreover, as this semantic enrichment focuses on concepts and entities rather than raw tokens, and compares concepts and entities contained in the user query with those contained in the documents, it

[18] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/

[19] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

allows us to model the relation between user query and relevant documents in a more fine-grained way.

6.1 Task description

The first part of our work involved extracting and disambiguating all named-entities (people, location, organisation, etc) in the texts. To do so we employed the Python toolkit spaCy and the English pre-trained model 'en_core_web_lg' (described as: 'English multi-task Convolutional Neural Network trained on OntoNotes, with GloVe vectors trained on Common Crawl'). Three scripts drive this part of the process, focusing on performance gains via multi-threading and extracting entity occurrence frequencies.

We have also explored how to link detected entities to Wikipedia articles. Whilst having a high potential to reduce ambiguity in website and search query content, most available pipelines have drawbacks: the web-service TagMe[20] for instance relies on an API, which is a bottleneck when processing 350,000 documents. On the other hand, spaCy does not offer pre-trained models for entity linking and DBpedia spotlight models[21] are not readily accessible. Furthermore, there might be issues with data protection and intellectual property rights, since we would have to disclose data to a third-party server like TagMe via an API. For this reason, we provide a shallow disambiguation of entities relying on outlink statistics derived from Wikipedia.[22]

6.2 Experimental set-up

Whilst many options for parallelising the tagging process are available, the overall speed with which the full dataset is tagged is limited by the rate at which the data can be read in and the results can be saved. Therefore, we used an optimisation process to reduce the time required for processing the whole dataset. We used the `ThreadPoolExecutor` from the `concurrent.futures`[23] Python library to parallelise the code, and then

[20] `https://sobigdata.d4science.org/web/tagme/tagme-help`
[21] `https://www.dbpedia-spotlight.org/`
[22] `https://github.com/fedenanni/Reimplementing-TagMe`
[23] `https://docs.python.org/3/library/concurrent.futures.html`

compared run times over a smaller dataset of 300 files, between fully sequential and parallel versions. The benchmark test was carried out on a 64 core Microsoft Azure[24] virtual machine with 32GB of memory and SSD storage.

Threads	Time (sec.)
1	99
5	49
7	50
10	78

Table 1: Benchmark times from processing files in parallel by number of threads.

As can be observed in Table 1, running 5 threads in parallel appears to be the optimal setting, beyond this point using more threads does not reduce processing time. Each text file is read as input and has a corresponding JavaScript Object Notation (JSON)[25] file output, which causes the execution to be input/output (I/O) bound. Benchmarking was not conducted beyond 10 threads, and we cannot adequately explain why execution time sharply rises with additional threads.

6.3 Results

We processed 357,831 files from Dataset1 and focused on files containing the words 'speech' and 'news' in the filename. Please refer to section 5.2 for details on why we focused on these files. The processing took place over night and we did not capture the exact run time. Using a virtual machine with fewer cores may seem appropriate, but Azure virtual machines scale core count, memory, network and I/O bandwidth together. As we were I/O bound, an Azure virtual machine of the same type with fewer cores would also have reduced I/O bandwidth, making for slower performance.

We placed the content of the obtained 357,831 JSON files into a SQLite[26] database table `DocumentEntity` defined as follows:

[24] `https://azure.microsoft.com/`
[25] `https://www.json.org/`
[26] `https://www.sqlite.org/`

```
CREATE TABLE IF NOT EXISTS
DocumentEntity
(
    [ent] TEXT, -- Entity
    [cat] TEXT, -- Category of Entity
    [doc] TEXT, -- Document name
    [year] INT  -- Year of the document
)
```

6.4 Reproducing results

The reader interested in reproducing the results obtained in this work can access the scripts available in the companion GitHub repository.[27] The script named `multithread_calc_entities_spacy.py` processes a list of file names as input and save the output entities as JSON files. An extra parameter sets the number of threads to be processed in parallel. In addition, a version with serial processing was made available (please refer to `calc_entities_spacy.py`).

6.5 Conclusions

Entities are an established way of exploring the collection of web archives [Ruest and Milligan, 2019]. We show that annotating a very large collection with out-of-the-box tools is achievable in a short amount of time and the output, which comprises names of political leaders (e.g. David Cameron, Theresa May), organisations and events (e.g. the Olympics Games, the European Union), enables a fine-grained exploration of the collection. Nevertheless, it is important to consider that named entity recognition and entity linking tools are still far from perfect, especially when applied in noisy contexts.

7 Experiment: document embedding

The second experiment focused on preparing for semantic search over the collection, to improve over simple string matching. Being able to

[27] `https://github.com/alan-turing-institute/DSG-TNA-UKGWA`

understand the information needs of a user and offering back the most relevant documents has always been the central goal of information retrieval approaches [Manning et al., 2008]. We have approached such a task by leveraging current research in distributional semantics using word embedddings [Mikolov et al., 2013] and pre-trained language model approaches [Devlin et al., 2018], and their application in search scenarios over large-scale collections [Mitra and Craswell, 2017].

7.1 Task description

Thinking of a user engaging with the UKGWA, this experiment focused on returning the most relevant document to a user query. A number of different approaches were used for this task.

One of the simplest approaches is keyword matching, currently used by the vast majority of web archives that provide a search service. Given a string entered by a user (which we will call a search string), keyword matching finds all the documents containing this search string. The advantages of this approach are ease of implementation, explainability, and speed of information retrieval. However, keyword matching does not consider the semantics of a user query. For example, if given the search string 'games', a user may refer to the 'Olympics Games' or computer games, the system does not know which.

A common solution to this problem involves document embeddings. In this approach, the user query, as well as all websites in the database, are identified with a vector of 300 dimensions, for instance. The website in the database, whose vector is most similar to the vector of the user query, is then returned as the most relevant answer to the user query.

7.2 Experimental set-up

In our experiment, we created document embeddings by modelling the entities extracted from the previous entity recognition and disambiguation experiment (see previous section). We used entities instead of words within the websites, to overcome the fact that the data contained noise from JavaScript code and HTML tags. It is important to mention that this was the only cleaning step made before creating the embeddings.

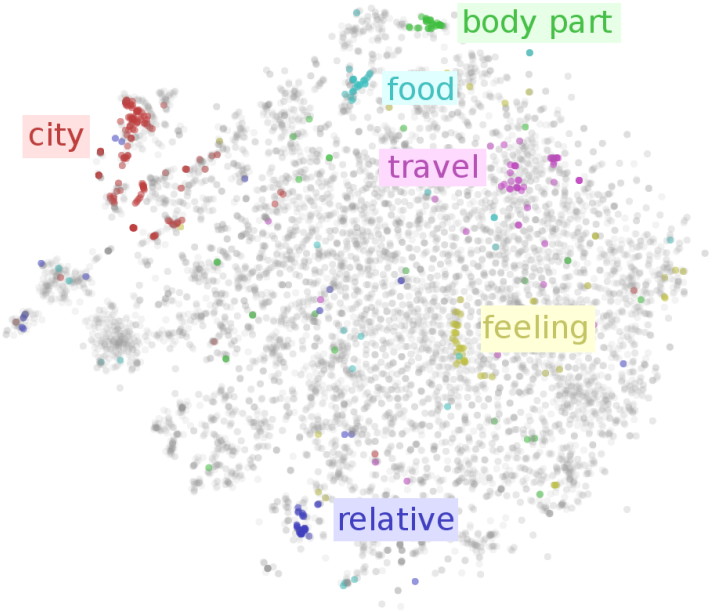

Figure 5: Illustration of document embeddings. Each point is a document embedding, representing a website from Wikipedia. This figure was not obtained from the datasets described in this report and features here for illustration purposes. Source: Christopher Olah, https://colah.github.io/

To create the document embeddings (see figure 5), i.e. the embeddings of each website page, we used the Paragraph Vector model, an unsupervised neural network approach. The Gensim Python library by Řehůřek and Sojka [2010] provides a ready-to-use implementation of this model (named Doc2Vec), which we have employed using default parameters.

The state-of-the-art algorithms for word embeddings (vector representations of words instead of documents), notably ELMo[28] and Bidirectional Encoder Representations from Transformers (BERT)[29], employ Recurrent Neural Networks (RNNs). These algorithms offer a higher quality of embedding, but are subject to substantially higher computation time. While we have experimented with them, we were not

[28] `https://allennlp.org/elmo`
[29] `https://arxiv.org/pdf/1810.04805.pdf`

able to apply these techniques at scale during the Data Study Group week. Nevertheless, we expect to see an improvement in the results when employing them in comparison with traditional document embeddings. ELMo and BERT are usually not used for document embeddings, and would need careful changes for the given challenge. An option might be to employ recently released BERT-based sentence embeddings [Reimers and Gurevych, 2019]. Representing the corpus as document or word embeddings would help solve some challenges of search and recommendation: when a user performs a search, return a list of documents which are the most similar to the query terms.

7.3 Results

The initial computation of document embeddings for the 357,831 considered websites took approximately 3 hours. The size of all document embeddings is 2.1GB. The size of the resulting deep learning model is approximately 100MB. The subsequent computation of an embedding for an arbitrary user query takes less than a second.

The quality of the clustering results is assessed in the following section, which describes the clusters within the resulting dataset.

7.4 Reproducing results

To reproduce the experiments, one must first obtain the entities, as outlined in the previous experiment on Entity Recognition. The resulting entities are stored as JSON files, one for each website. In the current set-up, the JSON files would be stored in the directory `/data/entities/results[number]/`, where `[number]` is a number between 1 and 10.

If the outputs are stored in a different location, the `doc2vec_helper.py` file has to be edited accordingly.

Then, the `doc2vec.py` file can be run. This creates the document embeddings in the directory `/data/doc2vec/`. The Doc2Vec algorithm proceeds in iterations (also known as epochs), in each of which a better model is derived. The algorithm finishes after a specified number of iterations, and saves the resulting model to

`/data/doc2vec/model_[version]`, where `[version]` is a numerical timestamp. In addition, all intermediate models are stored in the same directory. The overall number of iterations, as well as other hyperparameters, can be changed in the `doc2vec.py` file. Such changes, notably higher numbers of iterations, may improve the quality of document embeddings.

Lastly, the generated model is used to map each document to a small vector (of 300 entries in our experiments), the document embedding. These document embeddings are saved in the same directory as the model, as `/data/doc2vec/results_[version].csv`, where `[version]` is a numerical timestamp.

7.5 Conclusions

The document embeddings were generated from the list of document entities, rather than the plain text. We made this decision because of the noise in the document texts, such as mark-up tags and comments. Another possible approach to generating document embeddings would be to take all the content of the documents, to clean the content and then to produce document embeddings from the list of the tokens from all of the documents. A number of steps would be needed to clean the content, such as removing all English stopwords, all non-alphanumeric characters, etc.

Another limitation of this approach is the challenge of processing a user query that contains a word which does not appear in the word embedding list. Currently, this problem is unaddressed and the word is ignored; however, more advanced approaches would employ sub-word or character embeddings to address the issue [Athiwaratkun et al., 2018].

Document embeddings have the potential to improve search quality and to assist in clustering the dataset into themes, which can then be used for further analysis. The runtime of the Doc2Vec algorithm we used (both for training and inference) is good, and suitable for practical use. However, the quality of the resulting embeddings can be improved, as analysed in the next section.

We suggest the following areas of focus for future investigation:

- Use other input data, specifically cleaned website data, or operate on words instead of entities
- Change the parameters of Doc2Vec (such as number of iterations)
- Consider alternatives to Doc2Vec such as ELMo or BERT, as discussed above

In particular, BERT models for word representations are a promising avenue of exploration, likely to be highly adaptable to a corpus such as the UKGWA, which is both entity-rich and context-rich. The potential of this approach is further advanced through the relative ease with which the framework can apparently work in conjunction with other NLP pipelines. Sentence-based approaches aid the contextualisation that would bring more powerful insights into the UKGWA dataset and would deliver a valuable complementary system to the current keyword-based full text search, without necessitating a radically different user experience.

8 Experiment: clustering

The third experiment involved the use of semantic representations of the documents, in order to form a bottom-up, data-driven organisation of the collection through a structure of interconnected topics. To achieve this, we explored the adoption of clustering algorithms, employing the previously extracted entities and topics. The combination of these clusters with semantic search and disambiguated entities allow us to offer a large number of options to the final user for exploring topics and trends in the archive, as discussed in the following section.

8.1 Task descriptions

A key part of an archival search is the ability to group documents together in well-defined topics, which a user may use to both narrow down their search results and to investigate topic evolution over time. The document embeddings described above should provide a means to do so, through the application of unsupervised clustering algorithms. We investigated

the adoption of two well known and widely used algorithms, k-means and hierarchical agglomerative clustering; both of them are implemented in python as part of the Scikit-Learn[30] SciPy packages.

K-means works by randomly initialising a predetermined number of centroids (the centres of the clusters), then iteratively assigning points (here documents) to their nearest clusters and then recalculating the centre of these new clusters. This is computationally inexpensive to perform, but the resulting clusters may be of poor quality due to the lack of structure or relations between them. Hierarchical agglomerative clustering addresses this issue, by iteratively merging clusters, starting from a singleton (that is, clusters containing one document only).

8.2 Experimental set-up

To evaluate the quality of these clusters for a given number of topics, we apply silhouette analysis. This is a simple but effective metric: first, the mean distance of each point to all others within its cluster is calculated (call this $a(i)$ for point i). Next, define $b(i)$ to be the smallest mean distance of i to points in another cluster, and define the silhouette score $s(i) = (b(i) - a(i))/\max(a(i), b(i))$. This value lies between -1 and 1, and by taking the mean value over the full dataset we may quickly evaluate cluster quality. A value close to one corresponds to well clustered data, while values around zero demonstrates that clusters are overlapping. To reduce computational expense of calculation, these scores may be approximated via a random sample of the full dataset.

Finally, so that clusters are interpretable for users, we require a way of automatically producing labels for these groups, that may then be refined by domain experts if desired. To do so, we considered the tagged entities present in a number of documents closest to each calculated centroid. By exploring those most unique to each cluster, we can in theory list key entities for each group and thus interpret what topic it corresponds to (as done by Lauscher et al. [2016]).

[30] `https://scikit-learn.org/`

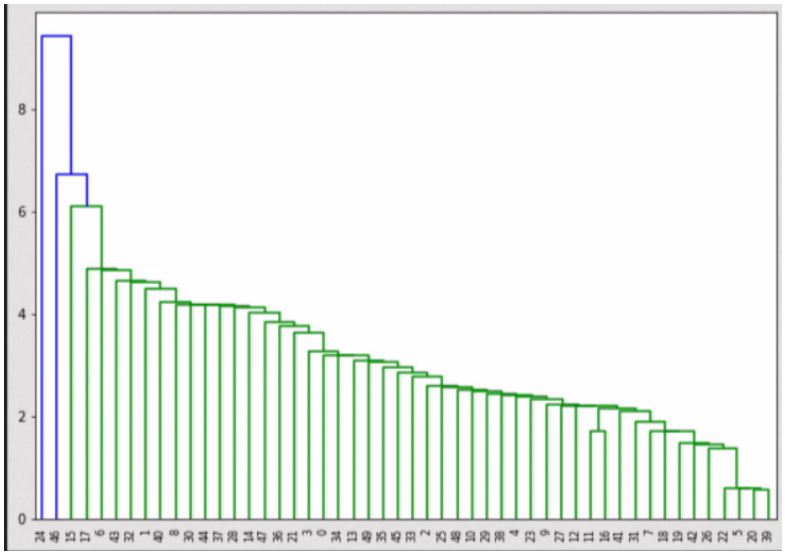

Figure 6: Hierarchical clustering on the document embeddings. Shown are only the top layers.

8.3 Results

The result of our initial experiment is the dendrogram shown in Figure 6. This is a diagram that shows the hierarchical relationship between objects, in our case document embeddings. While in this first attempt the document embeddings show no clearly defined clusters (i.e. a tree with distinct macro-groups), further work on document pre-processing, feature extraction and hyperparameter tuning (number of clusters, distance threshold, kernel k-means (see Dhillon et al. [2004])) would guide us in generating an interpretable structured overview of the collection.

The auto-labelling approach, attempted through taking 100 nearest documents in the embedding space to each centroid and then taking the most common entities of these documents is also limited by the large amount of boiler-plate content present in the datasets. Many documents (even when treated as embeddings) would have similar most frequent entities.

All the steps described in this section have been reached relying upon the scikit-learn Python library, in particular the clustering functions.

8.4 Conclusions

As discussed above, the presence of boilerplate text such as banners, links and disclaimers have a drastic negative impact when using the Doc2Vec model and consequently on the clustering of web pages.

Future work should therefore focus on two core issues of contemporary research on enhancing access to web archives:

- Improving content extraction from web pages, while in parallel
- Developing approaches to deal with highly noisy collections

Concerning the second point, approaches that leverage pairwise mutual information could be applied to identify the most relevant entities per cluster, ignoring the underlying noise.

Additional clustering approaches could be investigated. For instance, one could create a dissimilarity matrix of documents using Doc2Vec and cosine similarity, then treat the matrix as a network of documents and apply community detection methods to it. This works by thresholding the dissimilarity matrix, forming a weighted graph through MST-kNN (Minimal Spanning Tree (MST) and k-Nearest Neighbour (kNN), then permitting application of Markov stability approaches to community detection across a range of resolutions. This could also allow us to better understand nested modular structures inherent in the data (i.e. subtopics within a broader field). Other methods include manifold detection techniques to form the graph, such as contagion map.

9 Experiment: interface

This final experiment consisted of the exploration of different types of interface for the user of the UKGWA, based on the adoption of disambiguated entities, semantic search, and document clustering. The current interface offers a keyword based search, with some filtering options included to refine result sets. At the time of the event, the index covers nearly 400 million resources and contains a full-text search. However, this can lead to the results of a search delivering to the user too many hits, often in the order of thousands or millions. It is not transparent

in which order the results are ordered, and their relevance appears to the user be quite random at times. As explained in a previous section, this is indeed the case, but this is not currently explained to user.

9.1 Task descriptions

For this experiment, we explore the following questions:

- Is there a different way to visualise the data of the Web Archive?
- What interface would make it possible for the user to explore the data?
- How do we make this explainable for the user?

In order to make this visualisation possible a tool would have to be developed. This tool must be web-based, easy to implement on The National Archives web pages and be interactive. Additionally, end users should be given the opportunity to explore the data.

9.2 Experimental set-up

The interface should give the user an overview of the trends and topics within the web archive over time. To allow this, we would use the extracted entities and concepts to offer a comprehensive representation of the concepts in use at each point in time. Based on this temporal overview, the user should be able to drill down to specific sub-corpora that they are interested in. For this second part, the semantic search and clustering developed in earlier experiments would be used, as it would give the user a more refined search experience than the current full-text search.

The output from the previous experiments became the input for the interface experiment. Different types of graphs were trialled amongst the group to find the best way to visualise the temporal aspect of the data and their usefulness was assessed in collaboration with The National Archives experts present.

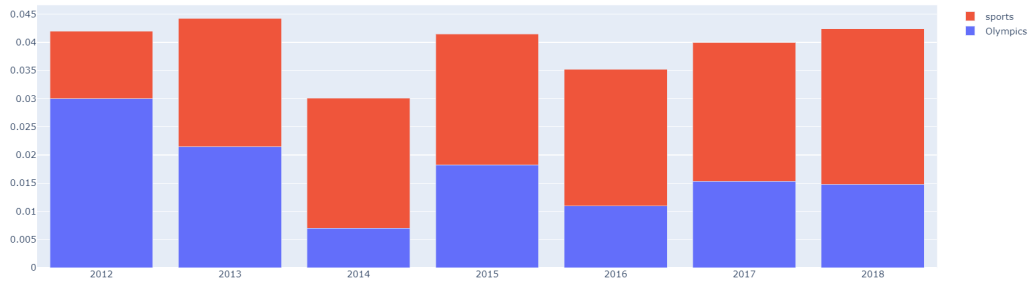

Figure 7: Initial prototype of the interface, presenting a comparison of the frequency of mentions of two entities over time.

9.3 Results

After trying a number of plots, we decided to use the stacked bar charts for the final implementation, as it easily allows the comparison of entities and concepts as topics over time. A visualisation of the graph with the input entities sport and Olympics can be seen in figure 7 (note: the entity Olympics also includes references to 'London 2012'):

Based on the interactive interface experiments, we have generated the following mock-up seen in figure 8.

We envisioned extra information accessible to the user by hovering over the question marks to make the results clearly understandable to a wide range of users. We also envisioned other graphs offered to the side, to allow data exploration in different ways. Finally, by selecting a sub-part of the results (for instance 'Olympics' in 2012), we would provide relevant documents based on the results of semantic search and clustering on this sub-part of the collection.

9.4 Conclusions

Next steps concerning the implementation of the interface would involve:

- The full integration of the developed components

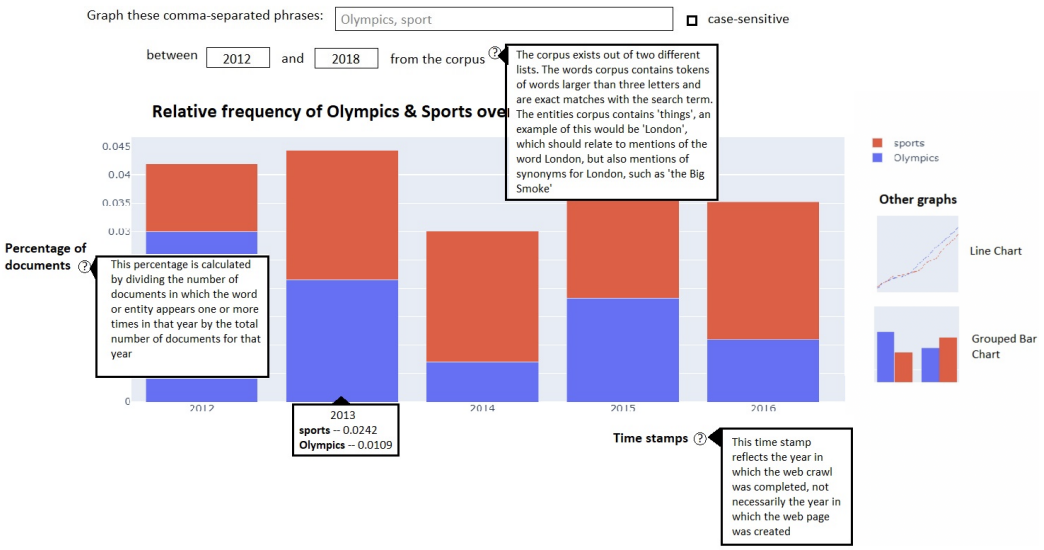

Figure 8: A mockup interface with improvements for the user, based on recommendations from the data owners.

- A user study on the pros and cons of integrating such novel interface in comparison with the previous string-matching search tool
- An analysis of the stability of such a tool when employed at scale, instead of only tested on a sub-sample of the collection

10 Future work and research avenues

Given further research and development, there is the clear potential for new search methods to enable users to discover content in the UKGWA, exploiting the rich information within.

This Data Study Group showed that the deployment of entity recognition and disambiguation at scale, using readily available software, is possible with data of this kind. Those methods would improve with additional contextual information, including that present in the metadata or catalogue and with data processed in a different way.

Document embeddings were shown to be useful to summarise the content in each document, by using those entities extracted. There would

be even greater successes, we expect, if these techniques were applied to all the textual content of each document, not just the smaller set of entities. If steps to improve the data by isolating or removing repeated web page content (e.g. navigation) were employed, then we are hopeful the the clustering techniques would prove to hold even further benefit.

Following best practice, for instance the ones described by The Turing Way Project: [The Turing Way Community et al., 2019], the technical and social design of any future system should be evaluated thoroughly before its interface and technologies are trialled with users.

In addition to scaling up and extending the work of the experiments outlined above, other areas of focus have been identified, as discussed below.

10.1 Expand corpus beyond HTML formats

HTML is the most common format in the UKGWA, but thousands of different MIME types[31], text and non-text, are also present. Extraction and processing of these document formats is somewhat different to the challenges we face when processing HTML, but they may make some issues such as duplicate detection easier.

There is untapped potential in the multimedia content that sits alongside the text-based content in the UKGWA. For example, speech-to-text conversions and/or image classification could be deployed to these formats and the extracted text passed through the same enrichment pipeline.

Backlinks (i.e. preserving and displaying to the user the location of a link to a file on a referring web page) will provide useful context and could aid more effective exploration around the topic. For example, the same page that provides a link to a report may also provide links to related documents.

[31] `https://en.wikipedia.org/wiki/Media_type`

10.2 Alternative datasets for HTML resources

A major issue, which was expected at the start of the week, was the heterogeneous nature of the data, in terms of variety of sources at disposal and the complexities of parsing their structure at scale. Using either the data from the source WARC files or the extracted text in the existing UKGWA Elastic index may reduce some of the data problems encountered.

If using HTML extracts in future, it should be determined whether certain HTML tags should be retained and preserved alongside the content as a way of encoding meaning information. However, retaining these risks introducing even more noise into the dataset through repetition of navigational and template content.

To overcome the noise in the dataset, we used entity recognition methods instead of operating on the words within the websites directly. This removed the boilerplate code effectively, but also some of the information content. The navigational features of web pages provide valuable contextual information, so reaching a balanced approach will take more evaluation.

10.3 Data and code in re-usable, open formats

The focus should not only be on systems, but on practice too, especially with regard to publishing datasets that third parties can access and process themselves. Several institutions and initiatives publish such data, which encourages research into their collections and raises their profiles in academic research circles. A good example is Common Crawl[32] and the permissive re-use terms attached to the vast majority of UKGWA content support a similar approach. Care would need to be taken to provide high quality documentation.

Similarly, the code used to generate datasets derived from the collection, along with all algorithms deployed on the collection, should be published under open re-use terms to aid transparency in developed systems and to support critical analysis, for example by forking.

[32] `https://commoncrawl.org/`

10.4 Combine interface with other visualisations

The user experience and sense of exploration could also be enhanced by providing complementary interactive services alongside the system proposed in this report. For example, including interactive network graphs could improve the user experience by placing their search terms in the context of the wider collection and allowing them to view the websites as entities possessing intrinsic properties.

The most obvious example of this is with a campaign website focussed on a particular issue; however, such an approach should give users the ability to explore topics and themes across multiple domains, as we know from anecdotal evidence that most users are not primarily interested in where something appears but rather when, why and in relation to what it appears.

Providing a temporal dimension to such a visualisation would also highlight one of the key parts of the archived web, namely changes over time in both structure and content.

10.5 Improve approaches for handling duplication

In order to avoid the skewing effect of repeated counting of the same entity in an unchanging document (URL), data could be pre-processed to mark up the first occurrence and the fact that it persists for a given period of time. As described earlier in this report, this approach depends on the ability to reliably identify and extract pertinent content in the first place, in order to assert that it has not materially changed between point X and point Y.

Generating scores between documents based on their degree of similarity may be a useful route into this problem. Such analysis is likely to be easier with non-HTML resources, such as PDFs, because, as static documents, they are less vulnerable to interference by changes to website templates, or other stylistic (rather than content-based) changes.

10.6 Improve integration with knowledge sources

Wikipedia, and its open linked data source, DBPedia [Auer et al., 2007], has been mentioned as a useful contextualising and enhancing service. Other third party sources may be equally promising for leveraging factual information. Additional datasets can be sourced from government to add structure to domain knowledge, by using, for example, ontologies of government structure, an early version of which is already available to The National Archives, derived from various sources of data, such as organograms[33] and data published as registers.[34]

Furthermore, consideration would be given to whether other third party data sources could be integrated to provide useful context. Examples include media organisations, data repositories, or other web archives.

10.7 Leverage existing search query data

Existing search data relating to the use of the UKGWA have enormous potential and future work should leverage these to identify potential focus areas. The existing UKGWA search service handles around 30,000 queries per month,[35] which offer a great basis for building comparative datasets for testing and validation and gaining insights into user experience and user success.

10.8 Track semantic change

In addition to the emergence of entities and themes over time, evolution of meanings is an avenue of future exploration. Recent research in NLP has developed various computational models for automatic semantic change detection (see overview in Tahmasebi et al. [2018]) and some studies have used web archive data (Basile and McGillivray [2018] and Tsakalidis et al. [2019]). This could be particularly interesting if additional sources of knowledge could be employed in the development of

[33] https://webarchive.nationalarchives.gov.uk/20120405144126/http://reference.data.gov.uk/doc/department/co

[34] https://www.registers.service.gov.uk/

[35] https://webarchive.nationalarchives.gov.uk/search/

computational systems, especially government-specific knowledge, but also third party sources such as other archives and media.

10.9 Develop customised tools for the communities

Building on existing tools by developing customised versions could be considered as an avenue of further exploration, and a useful way of opening new collaborations with the web archiving, digital preservation and broader research community.

One good example of such a set of tools is The Archives Unleashed Project.[36] The project received funding from the Andrew W. Mellon Foundation,[37] to develop a web archive search and data analysis tool to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web. One of the main aims of The Archives Unleashed Project is to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. The system is a suite of tools that uses WARC files as an input to provide historians, who do not typically possess the requisite skills for such data analysis, to interrogate the data and extract meaning from it.

10.10 Transparency

During the week the decision of The National Archives not to apply weightings to the search results was discussed. As a well-established method for improving discovery of resources, some positive and negative weightings may serve users well. Such a decision has hitherto not been taken due to the necessity of transparency: that is, to indicate clearly to users that a weighting has been applied, what it is and why. In implementation, we might choose to promote transparency and explainability of results by including an 'opt out' option for such weightings.

[36] `https://archivesunleashed.org/`
[37] `https://mellon.org/`

11 Team members

Fazl Barez is a PhD student at Edinburgh Center for Robotics. Fazl's main research interest lies in interpretability and safety. Fazl Facilitated this study group, helped with the analysis and report writing.

David Beavan is Senior Research Software Engineer – Digital Humanities in the Research Engineering Group in The Alan Turing Institute and Research Affiliate at the University of Edinburgh Centre for Data, Culture & Society. He has been working in the Digital Humanities (DH) for over 15 years, working collaboratively, applying cutting edge computational methods to explore new humanities challenges. He is currently Co-Investigator for the flag-ship Arts and Humanities Research Council (AHRC) funded project Living with Machines. David is co-organiser of the Humanities and Data Science Turing Interest Group, and is Research Engineering's challenge lead for Arts & Humanities in the Data Science for Science programme. He was Principle Investigator (PI) for this challenge.

Mark Bell is a member of the Research and Academic Engagement Team at The National Archives. Mark started his career at The National Archives working on the Arts and Humanities Research Council (AHRC) funded Traces through Time project, researching methods for probabilistically linking biographical records. He was also a researcher on the Engineering and Physical Sciences Research Council (ESPRC) funded ARCHANGEL project, which aimed to increase the sustainability of digital archives through the development of distributed ledger technology (DLT). Mark's other research interests include the use of handwritten text recognition to increase accessibility and enable new ways to explore digitised collections at scale, and the use of the UKGWA as a data source. He was Challenge Owner for this challenge.

John Fitzgerald is a PhD student at the University of Oxford, focusing on investigating how publication data can be used to understand the process of knowledge accumulation. By applying temporal network analysis and natural language processing techniques, the path of researchers through their career and their interactions can be explored at both a micro and macro level. This information can then be leveraged to make policy

recommendations, especially to developing countries, in how to improve the retention and development of the domestic knowledge base. He was one of the participants.

Eirini Goudarouli is the Head of Digital Research Programmes, and a member of the Research and Academic Engagement team at The National Archives. She has extensive experience working on digital and interdisciplinary research projects across the archival and higher education sectors. She was the Co-Investigator of the International Research Collaboration Network in Computational Archival Science (IRCN-CAS; 2019 - 2020),[38] funded by the Arts and Humanities Research Council (AHRC). She led discussions for this challenge, including the co-development of its research goals and the establishment of a formal research agreement between The National Archives and The Alan Turing Institute.

Konrad Kollnig is a PhD Student at the University of Oxford, studying data protection law. Previously, he read mathematics and computer science in Aachen (Germany), Edinburgh, and Oxford. He was one of the participants.

Barbara McGillivray is a Turing Research Fellow at The Alan Turing Institute and at the University of Cambridge. She is editor-in-chief of the Journal of Open Humanities Data[39] and Co-Investigator of the AHRC-funded project Living with Machines. She founded the Humanities and Data Science Turing Interest Group and her research interests are on computational models of language change. She has worked on semantic change detection from different sources, including the UK Web Archive JISC dataset 1996-2013. She conceived the idea of this challenge.

Federico Nanni is a Research Data Scientist at The Alan Turing Institute, working as part of the Research Engineering Group, and a visiting fellow at the School of Advanced Study, University of London. He completed a PhD in History of Technology and Digital Humanities at the University of Bologna focused on the use of web archives in historical research and has been a post-doc in Computational Social Science at the Data and Web Science Group of the University of Mannheim. He was one of the

[38] https://computationalarchives.net/

[39] https://openhumanitiesdata.metajnl.com/

participants.

Tariq Rashid is the founder of Digital Dynamics, which specialises in helping organisations understand the risks around their use of machine learning and automated decision making, covering issues such as data bias, algorithmic transparency, benchmarking and testing processes, and accountable governance. He is active in building communities around machine learning and artificial intelligence and leads Data Science Cornwall. He is also the author of an accessible guide to neural networks that has been translated into six languages. He is passionate about ethical and responsible use of AI and is part of an EU working group testing their emerging framework. He was one of the participants.

Sandro Sousa is a PhD student at the School of Mathematical Sciences, Queen Mary University of London. He is interested in emergent phenomena in complex systems, networks, segregation dynamics, urban transportation. His PhD focusses on quantifying the heterogeneity of spatially-embedded systems through random walks on graphs with particular interest on socio-economic segregation. Previously, he has worked as a Research Associate in a project funded by the Economic and Social Research Council and Fundação de Amparo à Pesquisa do Estado de São Paulo which looked into the relationship of spatial segregation and transport accessibility in London and São Paulo. He was one of the participants.

Tom Storrar is Web Archiving Service Owner at The National Archives, where he manages a team to deliver the UK Government Web Archive and the EU Exit Web Archive. In addition to delivering improvements to these online services, he is interested in developing innovative ways in which the archival data can be used for research. He has been involved in several projects involving the use of this data, exploring questions relating to its linguistic content, and network and longitudinal analysis. He was the Challenge Owner of this challenge.

Kirill Svetlov is a research fellow in quantitative finance at the Laboratory for Research of Social-Economic and Political Processes of Modern Society at Saint-Petersburg State University. His research interests are closely connected to stochastic process modelling and quantitative finance. His main mathematical interests are partial differential equations, inverse and ill posed problems, probability theory, stochastic processes,

scientific computing, and numerical analysis. He was one of the participants.

Leontien Talboom is a collaborative PhD student at The National Archives, UK and University College London. Her research is on the constraints faced by digital preservation practitioners when making born-digital material accessible. She has a Master's in Digital Archaeology and completed this with a dissertation focusing on NLP techniques to improve the discoverability of zooarchaeological material in unpublished archaeological reports. She was one of the participants.

Aude Vuilliomenet is a graduate from The Bartlett Center for Advanced Spatial Analysis at University College London. She is interested in using data science techniques and implementing IoT technologies to explore human interaction and nature dynamics in urban green spaces. She has previously employed NLP techniques to research food trends. She was one of the participants.

Pip Willcox is Head of Research at The National Archives and leads the Research and Academic Engagement team. She has worked at the intersection of digital scholarship and heritage collections for over 20 years, designing and delivering interdisciplinary research in fields including Web Science, digital musicology, history, and heritage collections. She is investigator of two AHRC-funded projects:[40] Engaging Crowds: Citizen Research and Heritage Data at Scale as Principal Investigator; and Deep Discoveries as Co-Investigator. She co-developed the goals of this challenge as part of The National Archives' ongoing collaborations with The Alan Turing Institute, and its research strategy.

12 Notes

The copyright and database right in material produced by staff of The National Archives under this collaboration is Crown copyright or Crown database. Free and flexible re-use of Crown Copyright material is granted

[40] https://www.gov.uk/government/news/government-investment-backs-museums-of-the-future

under the terms of the Open Government Licence[41] (OGL), and Crown Copyright and database right material is published under the OGL.

13 Acknowledgements

This work originated as part of the activities of the Humanities and Data Science Turing Interest Group.[42]

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and by the Living with Machines project[43] under the AHRC grant AH/S01179X/1.

The authors, participants and contributors would like to express their gratitude to the work of the following staff at The National Archives: Lynn Swyny, Copyright Manager; Howard Davies, Policy Manager; Jon Ryder-Oliver, Senior Business Partner; and John Sheridan, Digital Director. We acknowledge and thank the hard work of the following from the Alan Turing Institute: Jules Manser, Project Manager - Data Study Groups; Daisy Parry, Data Study Groups Administrator; Warwick Wood, IT Support Engineer; James Robinson, Senior Research Software Engineer; all our reviewers, and those in the Partnerships and Events teams who made the Data Study Group happen.

[41] https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

[42] https://www.turing.ac.uk/research/interest-groups/humanities-and-data-science/

[43] https://livingwithmachines.ac.uk/

References

Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*, 2018.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

Pierpaolo Basile and Barbara McGillivray. Exploiting the Web for Semantic Change Detection. In *Discovery Science 2018*, volume 5, pages 194–208. Springer International Publishing, 2018. ISBN 9783030017712. doi: 10.1007/978-3-030-01771-2_13.

Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 365–374, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014118.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

Max Kemman, Martijn Kleppe, and Stef Scagliola. Just google it - digital research practices of humanities scholars, 2013.

Anne Lauscher, Federico Nanni, Pablo Ruiz Fabo, and Simone Paolo Ponzetto. Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability. *IJCol-Italian journal of computational linguistics*, 2(2):67–88, 2016.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3045025.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.

Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. Building entity-centric event collections. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10. IEEE, 2017.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Nick Ruest and Ian Milligan. Web archives analysis at scale with the archives unleashed cloud, 2019. URL http://hdl.handle.net/10315/36119.

Amit Singhal. Introducing the knowledge graph: Things, not string. official blog of google, 2012.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change. In *Preprint at ArXiv 2018.*, 2018. URL https://arxiv.org/abs/1811.06278.

The Turing Way Community, Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, and Kirstie Whitaker. The turing way: A handbook for reproducible data science, March 2019. URL https://zenodo.org/record/3233853.

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. Mining the UK Web Archive for Semantic Change Detection. In *Proceedings of International Conference Recent Advances in Natural Language Processing*, 2019. URL https://www.aclweb.org/anthology/R19-1139.pdf.

turing.ac.uk
@turinginst