

Big Data for Law

Transforming the way we think about legislation

The challenge

Our system of law has evolved over 800 years. The Statute Book, the legislation Parliament has enacted or the government has made, tells many stories. However, comprising of 50 million words, with 100,000 words added or changed every month, it is simply too big and changes too quickly to easily comprehend as a whole system. However, by representing legislation as data and applying data analytics approaches, that changes. Finally, we can start to see the big picture.

The objectives

- Create a suite of tools to enable big data research with legislation through the provision of a Legislation Data Research Infrastructure (LDRI).
- Conduct user research to identify different researchers' needs of the LDRI.
- Develop new methods for representing, processing and querying legislation as data.
- Use the LDRI to find commonly occurring legal design patterns, researching a 'pattern language' for legislation.

Pattern language

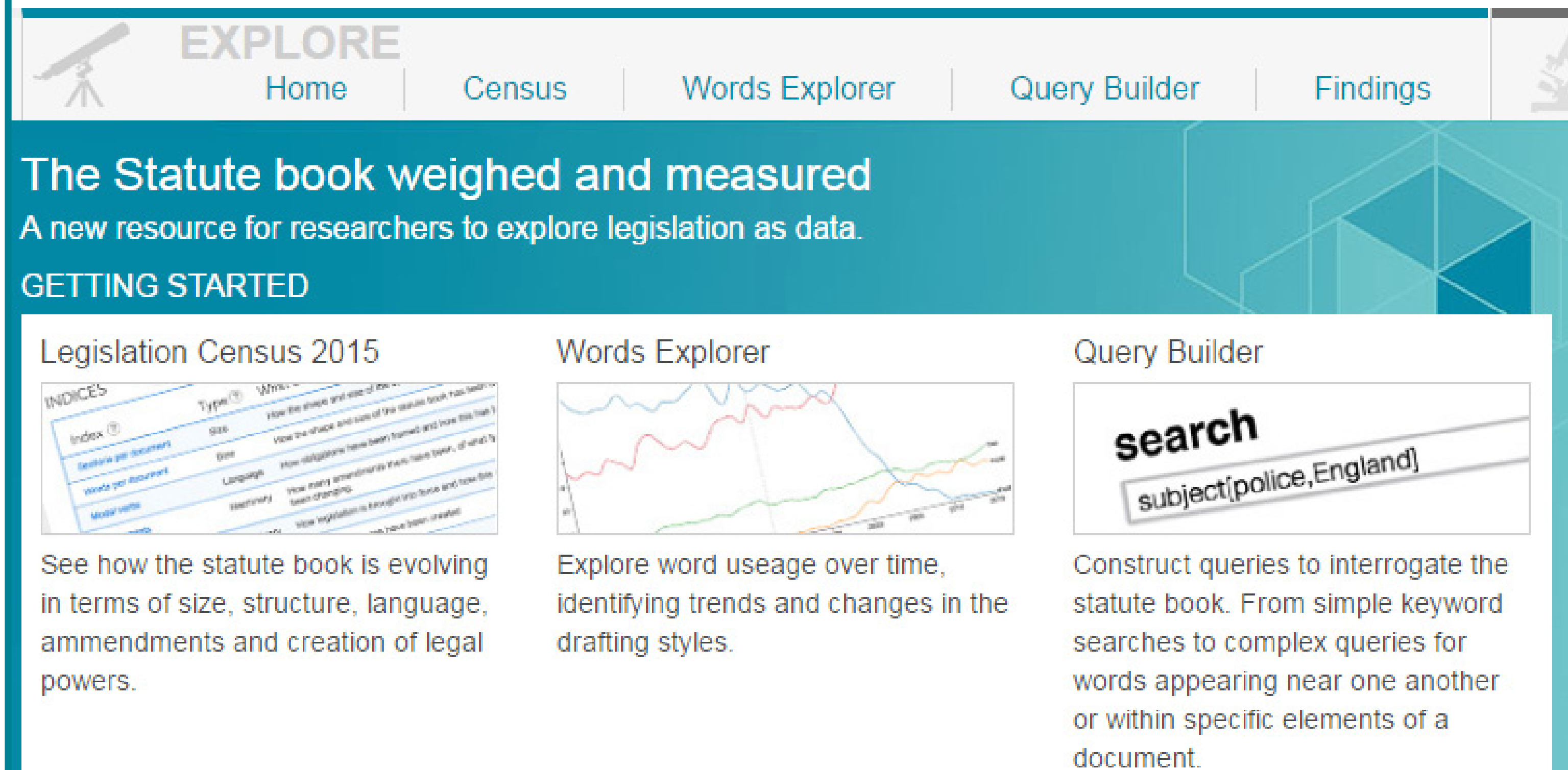
- We investigated whether there are commonly occurring legal design solutions to policy problems, in legislation, by exploring the concept of a pattern language.
- We worked with drafters of legislation to identify, name and specify over 20 patterns.
- We explored using Hohfeld's jural correlatives as a rigorous way to write the legal relationships each pattern instantiates.

Impact

- Our pattern language provided the catalyst and inspiration for the UK's four legislation drafting offices to produce their own catalogue of Common Legislative Solutions, to aid instructing Parliamentary Counsel.
- Our analytical tools are being used by government departments and The National Archives to aid preparations for EU Exit.

Authors: John Sheridan and Judith Riley

research.legislation.gov.uk



EXPLORE
Home | Census | Words Explorer | Query Builder | Findings

The Statute book weighed and measured

A new resource for researchers to explore legislation as data.

GETTING STARTED

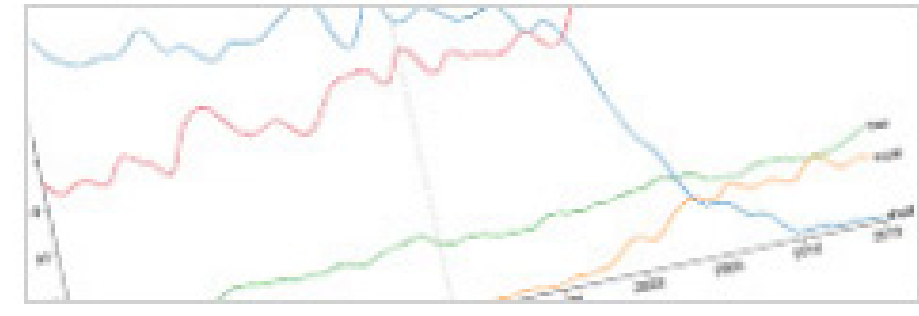
Legislation Census 2015

INDICES

Index	Type	What
Index 1	Size	How the size of the statute book has changed over time.
Index 2	Language	How the language of the statute book has changed over time.
Index 3	Structure	How the structure of the statute book has changed over time.
Index 4	Amendments	How the number of amendments to the statute book has changed over time.
Index 5	Creation	How the number of new laws created has changed over time.

See how the statute book is evolving in terms of size, structure, language, amendments and creation of legal powers.

Words Explorer



Explore word usage over time, identifying trends and changes in the drafting styles.

Query Builder

search
subject[police,England]

Construct queries to interrogate the statute book. From simple keyword searches to complex queries for words appearing near one another or within specific elements of a document.

The solution

- We developed a new website at <https://research.legislation.gov.uk> for researchers to use.
- We created new datasets, such as the first core reference dataset listing all UK legislation over 800 years, and an N-Grams dataset, counting the words and phrases.
- We developed an advanced tool that makes it easy to query legislation data, the Query Builder. This enables complex structurally and temporally aware searches as well as measurement of legislation as data.
- We have provided data-downloads of the Statute Book in a variety of formats.

Project lead: John Sheridan (The National Archives)
Co-investigators: David Howarth (University of Cambridge) and Professor Helen Xanthaki (Institute of Advanced Legal Studies UoL);
Advisory Board Chair: Sir Stephen Laws, KCB, QC, LL.D, Former First Parliamentary Counsel; K. Krasnow Waterman (CSAIL MIT)
Partners: Office of Parliamentary Counsel; Incorporated Counsel of Law Reporting; Lexis Nexis UK

Machine Learning

Digital transformation

The National Archives holds approximately 200km of paper documents in its repositories, spanning a timeframe of almost 1000 years. However, IBM estimate that 90% of the world's data was created in the last two years.

Where we are now

We are transitioning from a 30-year to a 20-year rule for receiving material from government departments. This means that we are still notionally in the mid-90s, when printing and filing was still the prevalent mode of preserving documents. The manual processes of appraisal, selection, and sensitivity review, remain just about practical.

Where we need to be

However, this is just a small wave arriving in advance of the digital tsunami when we move to the late 90s and new millennium. In order to appraise, select and conduct a sensitivity review on all of the records, we will need the assistance of machine learning to complement human reviewers. New technologies bring new opportunities too, enabling us to analyse and understand born-digital and digitised collections at scale like never before.

Digital skills

Machine Learning requires specialist skills, but it also needs to be understood by non-specialists.

Hackathon

On 7 to 8 December, we ran a machine learning hackathon for staff in our digital teams. An initial session introduced algorithms and techniques to demystify what is perceived to be a complex subject. Over 30 members of staff took part in the event, all were absolute beginners, many were non-technical.

Next steps

As well as learning new skills, the event helped to inspire ideas of how machine learning could be used in The National Archives' systems. Two ideas that we're now investigating further are automated recognition of coding languages and topic modelling catalogue descriptions.



Machine learning projects

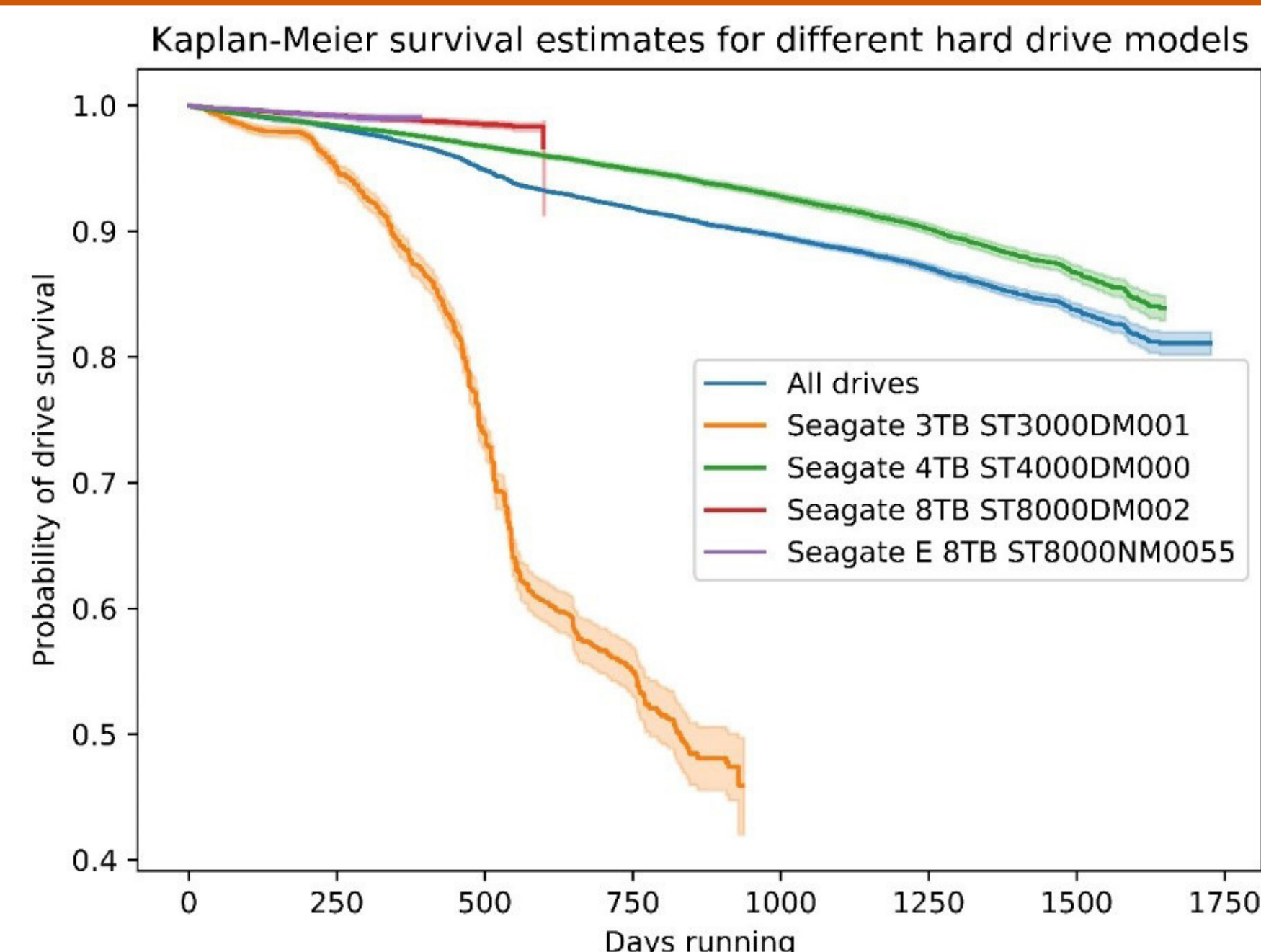
At The National Archives we are undertaking several different initiatives to explore the potential of machine learning technology:

- eDiscovery tools for appraisal and selection.
- Transkribus Handwritten Text Recognition.
- PhD projects in sensitivity review and understanding large scale web data.
- Cleaning crowdsourced data.

A Bayesian model of preservation risk

The nation's memory is at risk

Preservation risks to digital records arise from change. Preservation processes exist to reduce our risk exposure to an acceptable level. However, in the digital world, change is continual and our current qualitative approaches to risk assessment cannot provide us with adequate assurance. They are heavily based on human judgement which struggles to balance multiple, often interdependent, risk factors and can offer only a general indication of threat levels. As we evolve from an archive of paper into the disruptive digital archive of the future, we will require a quantitative, approach to understanding and managing digital preservation risk.



All storage media are susceptible to failure, putting digital objects at risk of loss or damage. The model will be driven by data, including hard drive survival estimates (data from Backblaze, plotted with Lifelines in Python).

A new approach to risk

We need a new approach to measuring and managing risks to the digital archive:

- Grounded in data but flexible enough to accommodate our changing understanding.
- Encompassing a wide range of threat factors, looking beyond format diversity to system dependencies, software, technical skills and organisational policies.

The disruptive digital archive

We will trial a statistical approach to quantify our exposure to digital preservation risk. A Bayesian approach is well suited to this problem. It will enable us to operate in the face of complex and imperfect information, and offer the potential to refine and adjust the model over time to reflect new data, the shifting environment and our changing risk appetite.

A model will allow us to simulate different scenarios, working through various options to explore the impact of potential risk reduction strategies. This will, in turn, help us determine how best to invest resources (both human and financial) to combat threats to the availability, identity, authenticity, persistence, renderability and understandability of the digital archive.



In this example of file corruption, part of the bit-stream has been lost (set to null bytes), leaving the record unrenderable. The risk model will let us explore the probability, causes and most effective mitigations of such preservation failures and will help us prioritise actions to mitigate different types of preservation risk.

Expected outcomes

- A predictive model, driven by an evolving understanding of rates of failure and change.
- A data driven approach to measuring the impact and cost of risk interventions.
- Better targeting of scarce resources to achieve optimal on-going threat reduction.
- Exploration of specific threat scenarios to test and validate the model.

Transformation of The Gazette

The Gazette is the UK's official public record. It publishes notices, such as bills receiving royal assent or the incorporation of a company, datasets and special supplements.

A brief history:

Before 1996

- Content was typeset and the process print-driven.
- Approximately 250,000 copies of London, Edinburgh and Belfast Gazettes were printed daily.

Since 1966

- 1998: information started to be captured natively in XML.
- 2014: a rebranded and technically advanced website was launched.
- Users are able to search the highly structured content for specific categories and accurate free text searches.
- 2018: just 80 copies are printed daily.

Submitting a notice

Gazette notices are submitted as full text or as data, where the only variable elements are provided. Data is merged with a template to produce a full notice.



When published, the RDFa is extracted from the notice and stored as RDF in a triple store. This provides both document views of the notice and data views of the notice.



The linked data application programming interface (API) allows a notice to be retrieved via format-specific URLs. The formats that can be queried include PDF, TTL, RDF, JSON, HTML and XML.



Longitudinal data sets

The Gazette's longitudinal data includes insolvencies, wills and probate, appointments and honours. Datasets go back as far as 1900, and the ability to find information since then has greatly improved.

Data submitted prior to 1997 was extracted from historic Gazette documents in the form of issues, supplements and indexes. The archive indexes were processed from 1900 onwards, and notices from 1998 onwards.

Valid, authentic data

Using extensible mark-up language (XML), digital signatures have been added to the website. The signatures check that the information is correct and unmodified, confirming the validity and authenticity of a digital document.

Data for commercial use

Gazette data is invaluable to organisations such as credit reference agencies, banks and government departments, who use it to support risk and opportunity management.

Companies can choose to have data delivered in a number of different formats, including XLS, XML and CSV. Insolvency, deceased estates and companies notice data are offered as part of the service.

See more at thegazette.co.uk/dataservice

Traces through time

Connecting people across the centuries

Imagine being able to enter a name and, with one click, find related documents from the millions of records held at The National Archives and beyond. This was the vision of Traces through Time.

The project applied data science techniques to connect records spanning centuries of history, linking the lives of ordinary people who appear in the archive. Our research questions were:

- How can we create fuzzy-linking approaches for historical data that is messy, incomplete and inconsistent?
- Can the linker accommodate diverse records of different types and from different time periods?
- Can we calculate a robust confidence score for each link to help users make informed choices?

Lane, Gerald Nassau Stewart
Reference: ADM 273/18/97
Description: Page 97: Gerald Nassau Stewart Lane.
Date: 1906-1918
Held by: The National Archives, Kew
Legal status: Public Record(s)
Language: English
Closure status: Open Document, Open Description
Access conditions: Closed For 75 years

Other possible matches

The following records may contain information about the person described above. As the links are found by computer analysis, we cannot guarantee they are the same individual or that every record in which the person is listed will be found.

Lane, Gerald Nasseau Stuart
Date of birth: 29 August 1899
AIR 76/286/20
● Air Officer's service record

Strong match

Lano, G N S
AIR 76/288/41
● Air Officer's service record

Weak match

The project

- Funded by: Arts and Humanities Research Council and The National Archives.
- Project partners: Institute of Historical Research, University of Brighton and Leiden University.
- Timeframe: 2013 - 2016
- Aims: Deliver a range of high-quality, tangible outputs that will help us link and contextualise our records to support decision making and improve access and navigation.

Outcomes and impact

- Algorithms and a statistical model for probabilistic linking of messy historical data.
- A linker and viewer to process and connect millions of records and evaluate results.
- User research to understand how to communicate and present probabilistic links.
- A data model and schema encompassing everything the records tell us about each 'person' and their connections.
- Over 600,000 links published via Discovery.

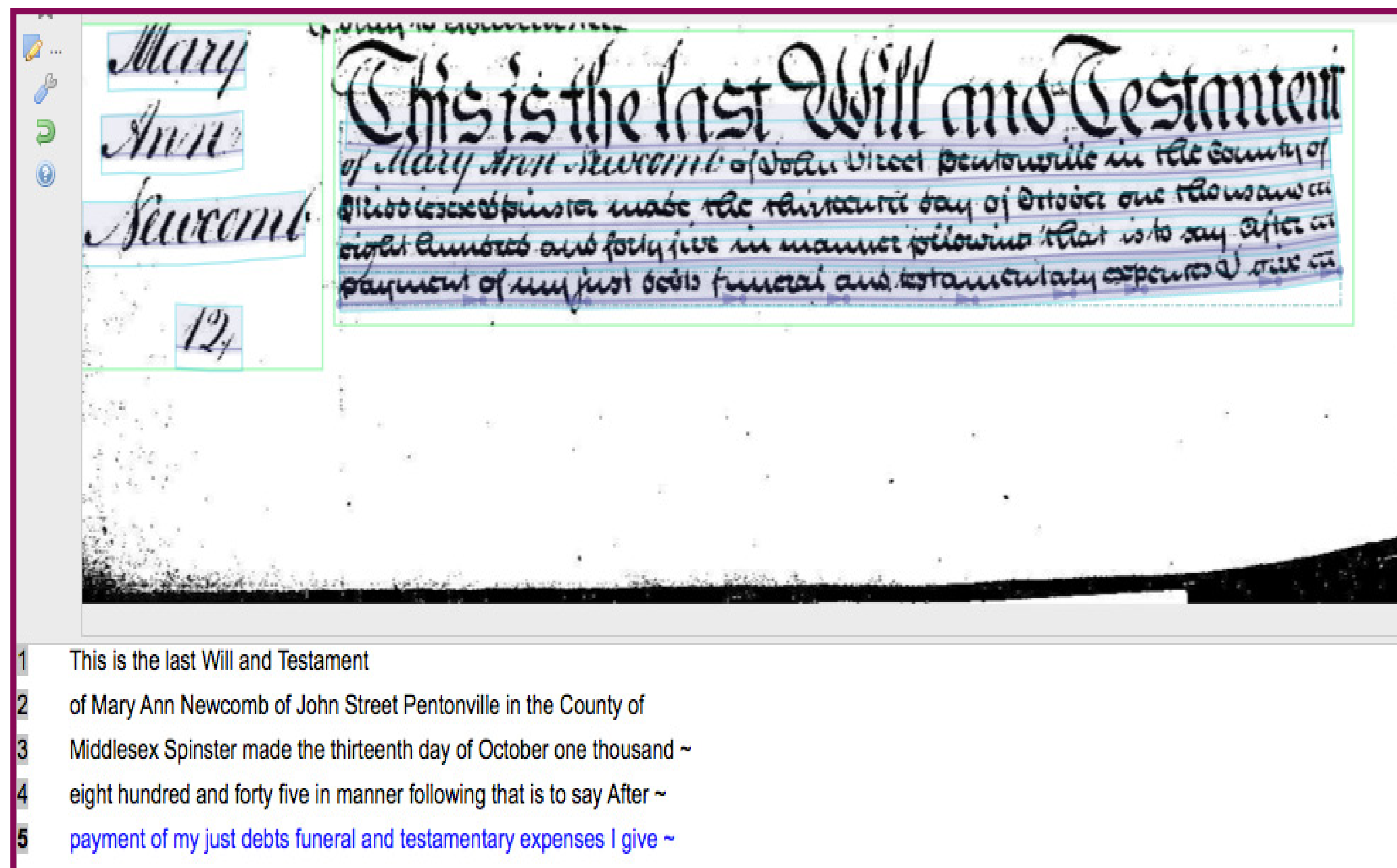
Can machines read handwriting?

The challenge

Optical Character Recognition (OCR) has radically changed how many researchers use historical texts and facilitated the application of new techniques such as sentiment analysis and named entity extraction.

However, Handwritten Text Recognition (HTR) is a far bigger challenge than OCR. OCR has only worked on printed documents, leaving vast swathes of the archive untouched by this element of the “digital turn”.

The National Archives has been experimenting with the Transkribus platform, built as part of the READ project, using our collection of wills (PROB 11).



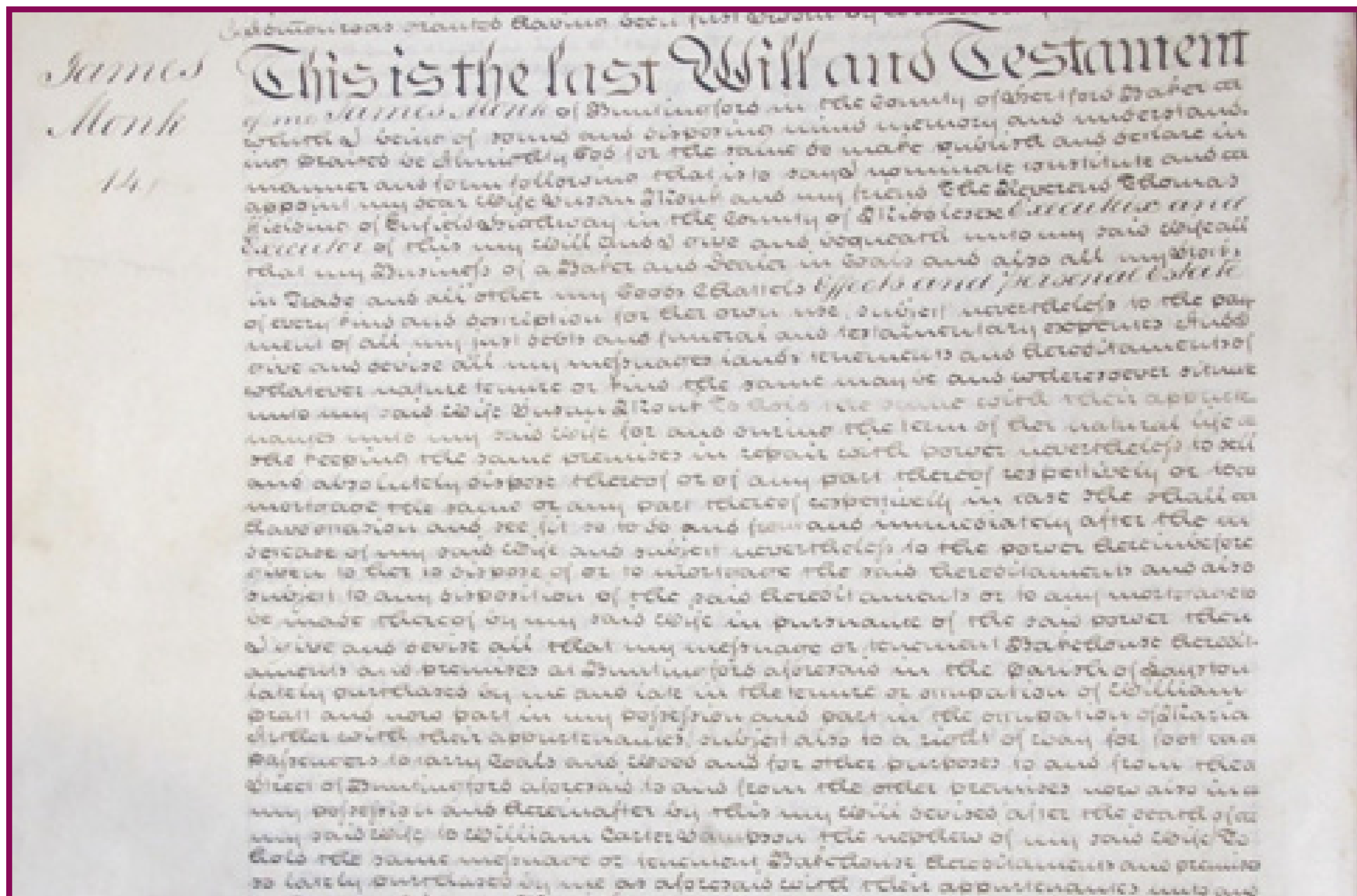
The process

Transkribus trains a recurrent neural network on a set of ground truth. Thus the model produced is trained on similar language and handwriting styles to that you want it to work upon.

We have been using our collection of PROB 11 wills due to their structured language and relatively uniform hand. Segmentation has proved a further challenge – especially where authors break with the traditional page structure.

In both segmentation and automatic transcription the process has been automated, but often requires some manual intervention.

Authors: Richard Dunley



The results

We started off by training a model on a relatively small set of ground truth (c.15,000 words). We achieved reasonable results from this with a word error rate (WER) of 39% and a character error rate (CER) of 21%.

Encouraged by this we trained a second model with 37,000 words of ground truth and have seen a marked improvement, with a WER of 28% and a CER of 14%.

A third model is currently being trained based upon a ground truth of 100,000 words and we hope to get a CER of under 10% which should be good enough to begin to work meaningfully with the outputs.

The next steps

The pilot project has been a great success in showing that HTR can work on difficult handwritten texts. The key questions for The National Archives going forward are:

1. How can we make the process less labour intensive?
2. What are the best ways to exploit the output – utilising its strengths in terms of scale whilst acknowledging it will not be perfectly accurate?

Using blockchain to engender trust

The challenges

The digital age presents new challenges to archives for safeguarding the data entrusted to them. Digital public records are intangible and so vulnerable to alteration, which could conceivably be carried out without detection and at scale. However, sometimes records must be modified to ensure they can still be used, for example to migrate their content as file formats evolve or become obsolete. This presented several research questions:

- How can we demonstrate that the record you see today is the same record that was entrusted to the archive twenty years previously?
- How do we prove that the only changes made to it were legitimate and have not affected the content?
- How do we ensure that citizens continue to see archives as trusted custodians of the digital public record?

Deep learning for Image Retrieval

Use deep neural networks trained on example data to extract digital signatures from whole images

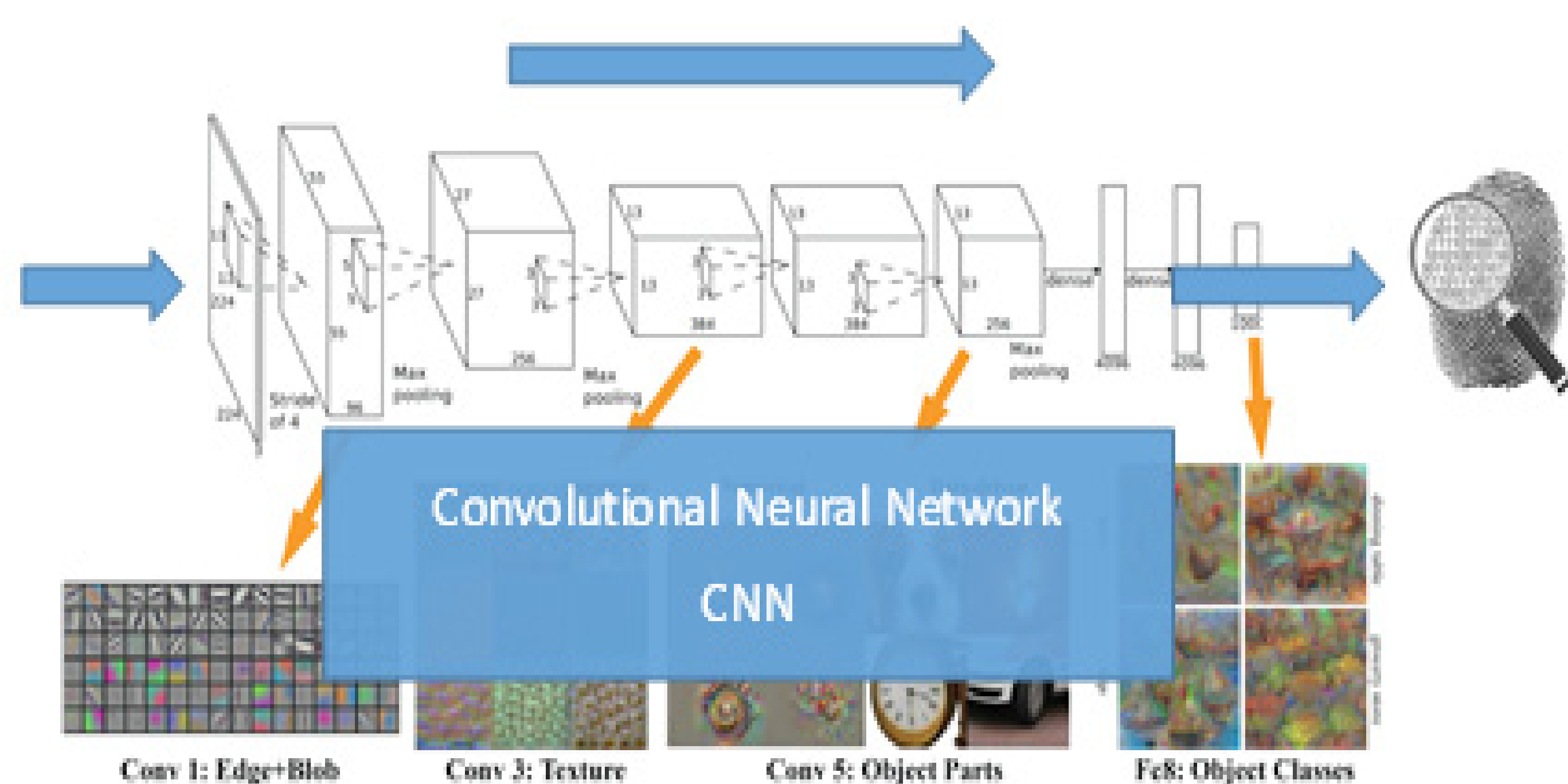


Image credit: John Collomosse, University of Surrey

The project

Title: ARCHANGEL - Trusted Archives of Digital Public Records

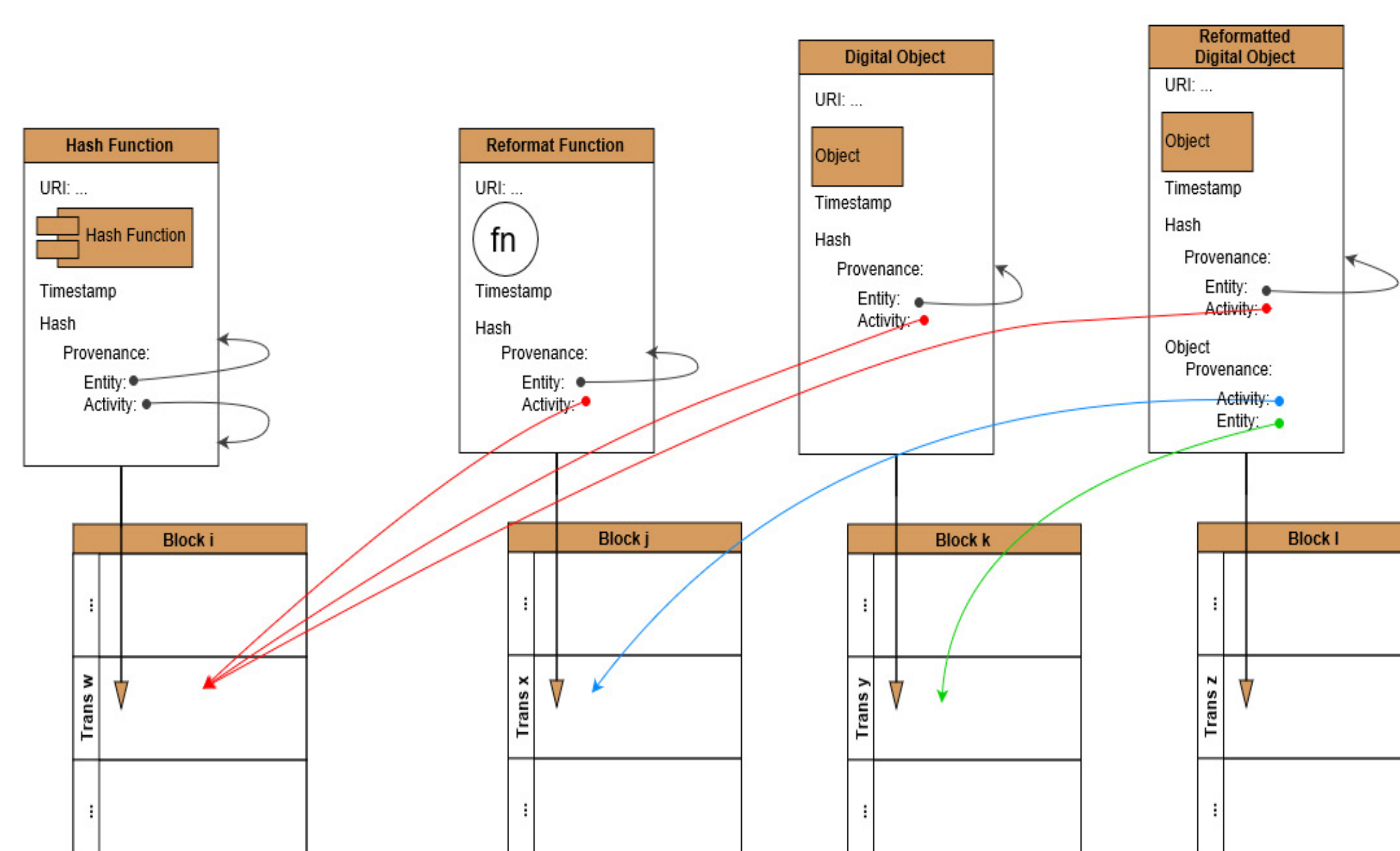
Funding body: EPSRC

Amount of funding: £487,428

Duration of the project: Apr 2017-Sep 2018

Project leads at The National Archives: John Sheridan, Mark Bell and Alex Green

Capturing Provenance in Blockchain



The research

The ARCHANGEL research project is exploring the potential of distributed ledger technology, such as blockchain, as a tool to maintain trust in digital records. We envisage that archives will be able to generate and register hashes of documents (like unique digital signatures) into a permissioned blockchain.

As the approach matures, the ledger would be maintained collaboratively and distributed across many participating archives across international borders as a promise that no individual institution could attempt to rewrite history. Where the record has been legitimately changed, hashes of the content alongside hashes of the code used to make the change, can also be registered on the blockchain.

The blockchain is immutable and distributed. The approach we will describe will result in the creation of many copies of a persistent and unchangeable record of the state of a document. This record will be verifiable using the same cryptographic algorithms many years into the future and therefore ensures the record's integrity.

This technology could transform the sustainability of digital public archives, enabling archives to share the stewardship of the records and, by sharing, guarantee the integrity of the records they hold.

Understanding users' mental models of digital archives

Challenges and opportunities

Discovery holds over 32 million descriptions of records held by both The National Archives and more than 2,500 archives across the country. Since its launch in April 2011, Discovery has evolved to incorporate features seeking to open up archives to wide-ranging audience groups, in addition to its original role as The National Archives' catalogue.

However, combining traditional archival methods with the opportunities provided by digital is causing great complication to users' understanding of how to use archival catalogues.

Applied user experience research methodologies

From September to December 2016, we focused on answering the question 'what is the Discovery user's mental model' by gathering user feedback through six user experience research methods:

- Data analytics- detailing a sample of 594,000 user sessions.
- Survey- gathering the feedback of 463 users.
- Details page widget- capturing feedback about the details page of 1,352 users.
- User interviews- providing insights into how users conceptualise and interact with Discovery.
- Workshops- observing users of varying levels of research skills and experience conducting complex searches.



Empowering users

Engaging with users identified two issues we need to overcome to enable users to successfully interact with Discovery:

1. Communicating complexity

The usability issues observed could be linked to the difficulty users have aligning their mental models with what is happening beneath the surface of Discovery.

2. Learning through use

The most experienced users were those that are able to build a knowledge of how to use Discovery through trial and error.

Future possibilities

This project identified the need for a comprehensive understanding of the mental models of users of archival catalogues. This will inform how future opportunities to enhance user experience in Discovery are identified.

Questions to consider for the future:

- How will the inclusion of born digital records redefine user expectations of digital catalogues?
- How can archives create engaging and informative digital experiences for users not familiar with traditional catalogues?
- How can we reimagine the archival catalogue to enhance user experience?

Webrecorder: archiving the un-archiveable

UK Government Web Archive

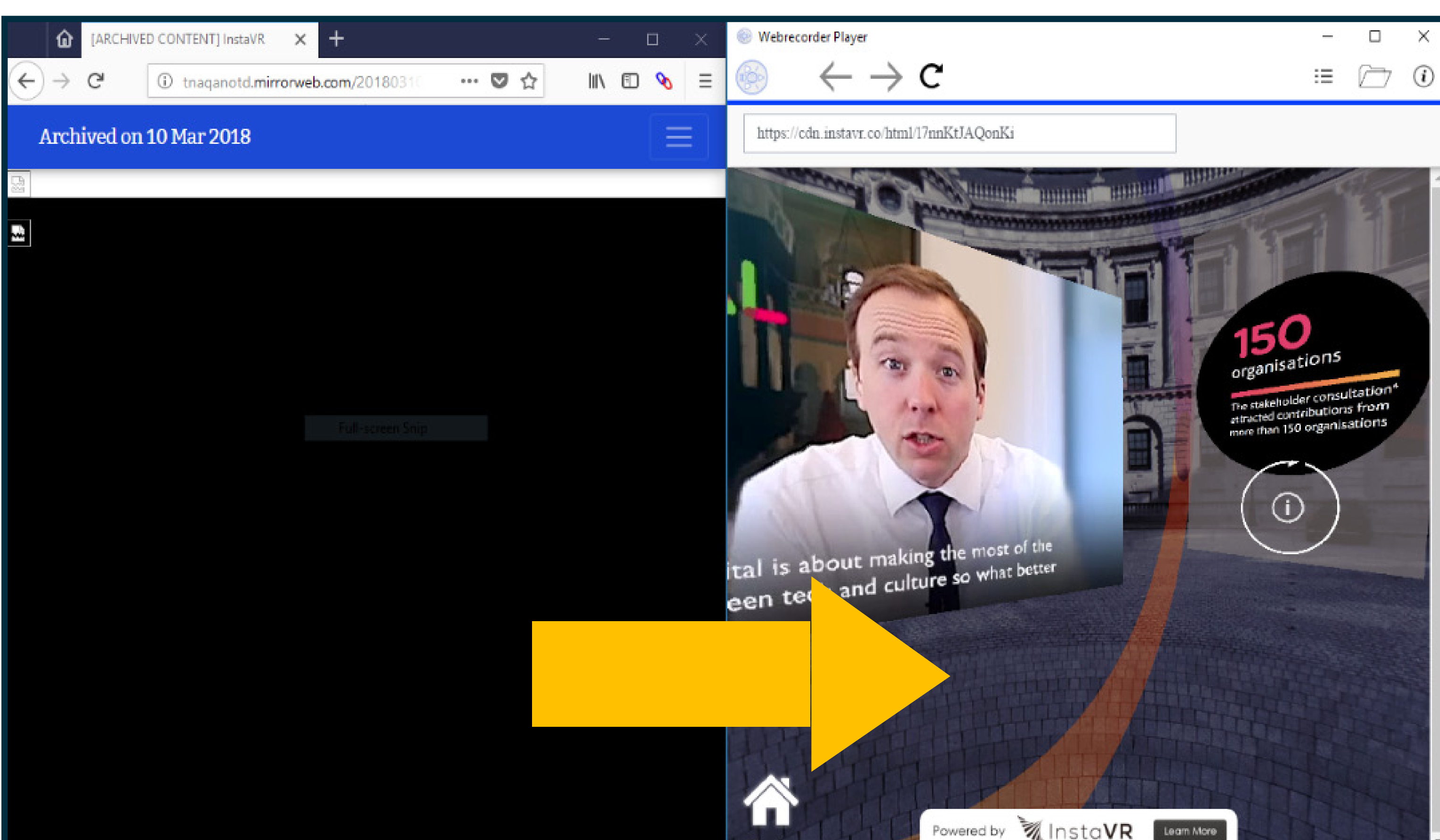
The UK Government Web Archive captures, preserves and makes accessible UK central government information published on the web. The web archive includes websites and social media from 1996 to present.

The challenge

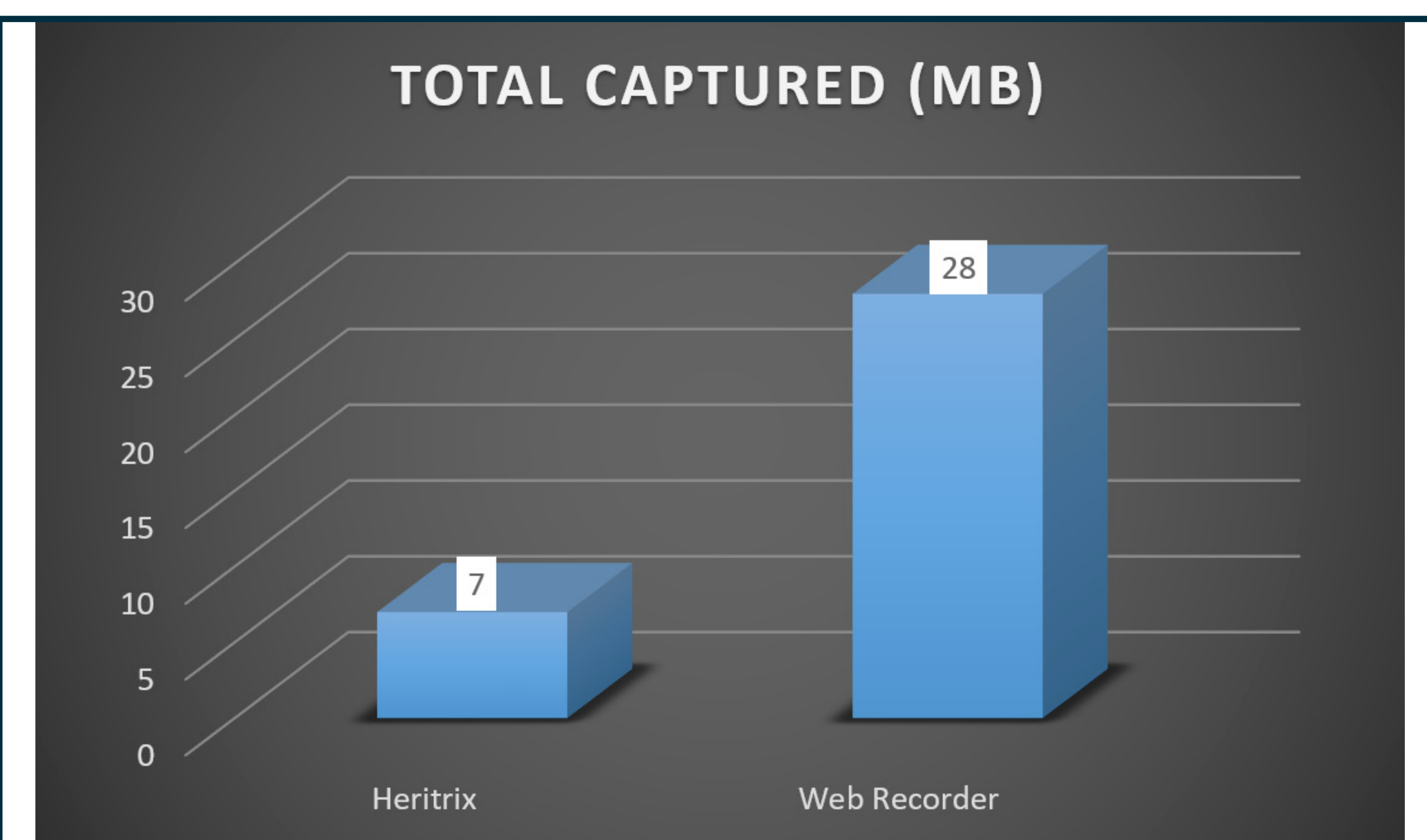
The established web archiving tool, Heritrix, is great at capturing traditional web pages but has limitations when it comes to capturing the latest web technologies. We now need to capture complex web pages with interactive content such as quizzes, timelines, animations and collaboration tools.

The solution

We used Webrecorder in conjunction with PYWB. Both are open-source tools developed by Rhizome to capture and replay content which caused problems for our established tools. We were successful in capturing and replaying the content of two very interactive sites.



DCMS (<https://cdn.instavr.co>)



Benefits

Webrecorder offers several advantages as a tool for web archiving:

1. makes archiving as simple as browsing.
2. produces ISO-complaint WARC files.
3. develops team skills.
4. augments established web archiving technology.
5. allows us to capture content quickly.
6. enables us to archive content that was impossible to capture.

Limitations

However, it's not a 'silver bullet' because:

1. it is highly manual and labour-intensive.
2. it is not yet scalable for large websites.
3. the operator (archivist) needs good knowledge of site being captured to avoid risk of human error.
4. it is not integrated with our regular web archiving pipeline.