Stories, sources and new formats:

the challenges of digitising large archive source collections

In the last two decades or so there has been a movement towards digitising large collections of original sources. These projects have had a range of purposes, approaches and target audiences but there can be little doubt that they have had a profound impact on the practice of history in the academic, educational and public domains. In this article **Ben Walsh** and **Andrew Payne** attempt to survey the broad landscape of the scale and impact of digitisation and assess how the digital landscape has affected the historical landscape.

The digitisation of history

The development of digital technologies has opened up tremendous new possibilities for the study of history. Digital imaging and digital storage have enabled the creation of vast digital libraries enabling academic historians, educators and the general public relatively easy access to very large collections of original documents and other sources from archive collections. The process began in the 1990s with the digitisation of document collections on to CD-ROM. One of the earliest instances of this was The British Library's series of CD-ROMs Sources in History. Since then digital imaging and digital storage have developed spectacularly such that high-resolution copies of documents, maps, cartoons, census and other official records and countless other types of documents are now available to users in universities, schools and homes.

And the digital collection continues to grow. Last year The National Archives delivered more than 640,000 documents to readers in their reading rooms at Kew and over 200 million downloads via their website and licensed associate partnerships. Clearly digitisation offers archives a powerful way to open up their collections to academic researchers, journalists and family historians as well as history teachers and students. Indeed, in many ways the internet was made for archives, allowing users to search collections and download copies of documents from anywhere in the world. The National Archives has transformed its online catalogue Discovery into the single point of access to 32 million descriptions of records; 22 million of these are in their own collection and over 10 million records in 2,500 other archives across the country. The vast majority of its collection, however, remains available only for physical inspection in the reading rooms at Kew, with around 7% of records having been digitised for online access.

Digitisation has been driven by four key factors: revenue generation, preservation, access and digital re-unification.

Revenue generation has become an important motive as archives have sought to widen their funding base. Name-rich collections offer an attractive option with a growing audience interested in family history and genealogy. By partnering with commercial organisations such as Ancestry.co.uk and FindMyPast.co.uk, it has been possible for The National Archives to provide much wider access to the census, military service and migration records to researchers around the world. At the same time the licensing agreements have generated income to allow for further investment in access and services. Digitisation also allows fragile and damaged records, such as the 'Burnt Collection' of First World War service records, to be preserved through reduced handling. Finally digitisation has allowed for the bringing together of physically separated content to create comprehensive online collections which allow for 'Big Data' projects such as the University of Sussex project on the household expenditure surveys of the early twentieth century which are held at Bangor University and The National Archives.

How is the process of digitisation contributing to academic scholarship?

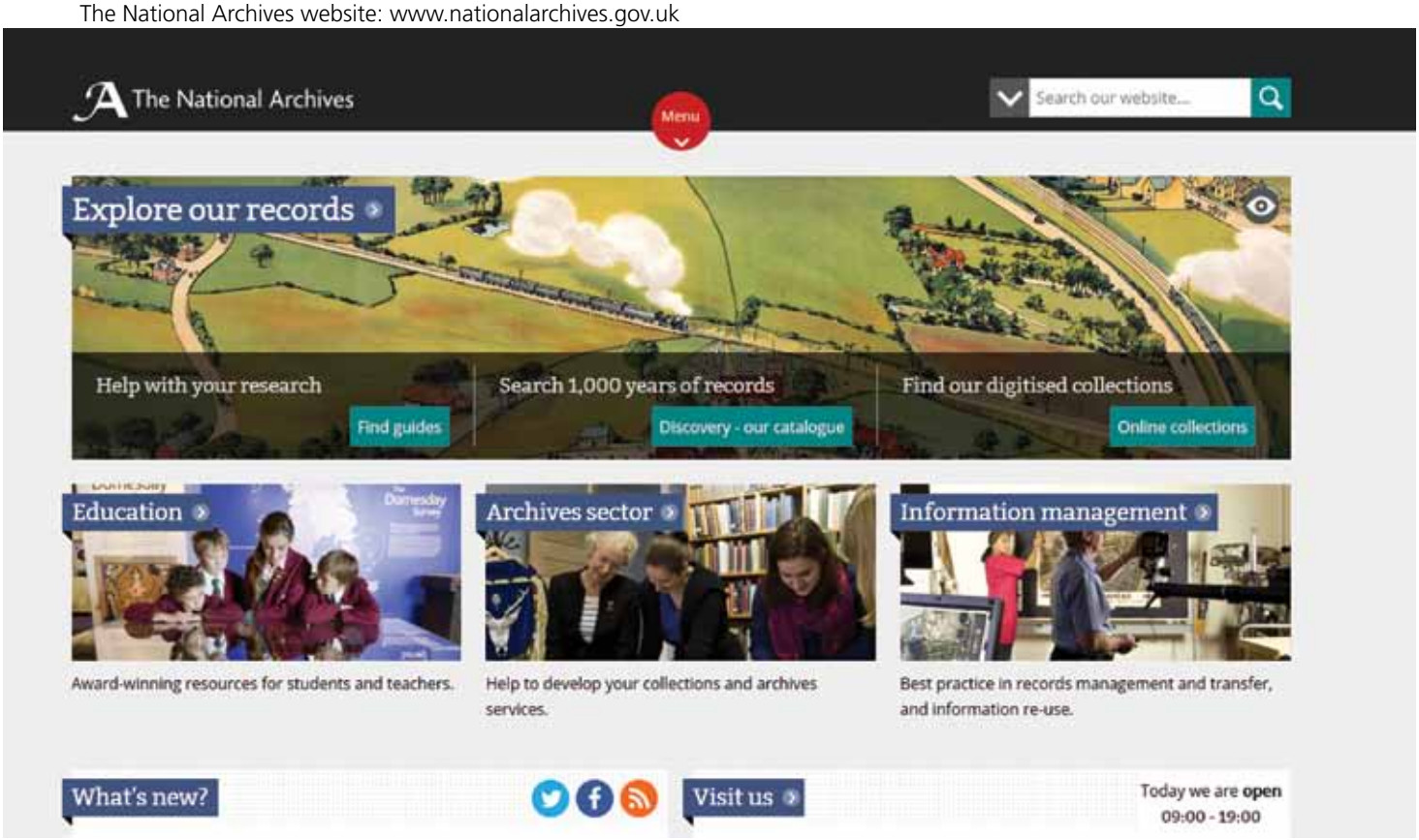
In addition to providing access to collections of source material which might be otherwise beyond reach, digital technologies allow historians to conduct new types of research. Digital resources offer up tremendous possibilities for searching, sorting, categorising and searching for patterns.

The British Living Standards Project

Professor Ian Gazeley of the University of Sussex has described how digitisation has transformed his research practice, particularly in the British Living Standards Project, based on British government household surveys:[1]

> The digitalisation of twentieth-century British household expenditure records has allowed

The National Archives website: www.nationalarchives.gov.uk



me to tackle bigger questions than was possible hitherto. I've used the results of household budgets studies in my research since my PhD thesis 30 years ago, but these were all small-scale enquiries that did not present significant issues with respect to capturing the information they recorded. The 1953/4 Household Expenditure Survey is immense by comparison and was designed to be a stratified random sample of all British households. It is quite rare to find a large survey where all the original records survived. Capturing the expenditure information recorded on a daily basis for three weeks by all individuals in 13,000 households would not have been feasible without modern digitisation techniques, as there are about one million hand-written individual daily expenditure records in the survey. We have digitised the material to be in keeping with the original records as faithfully as possible. The data we have extracted from this survey has allowed us to answer a number of questions relating to the standard of living in Britain at mid-twentieth century, and now that this data is in the public domain, I'm sure that in the future it will be used to address questions that we did not foresee at the point of digitisation.

England's medieval immigrants

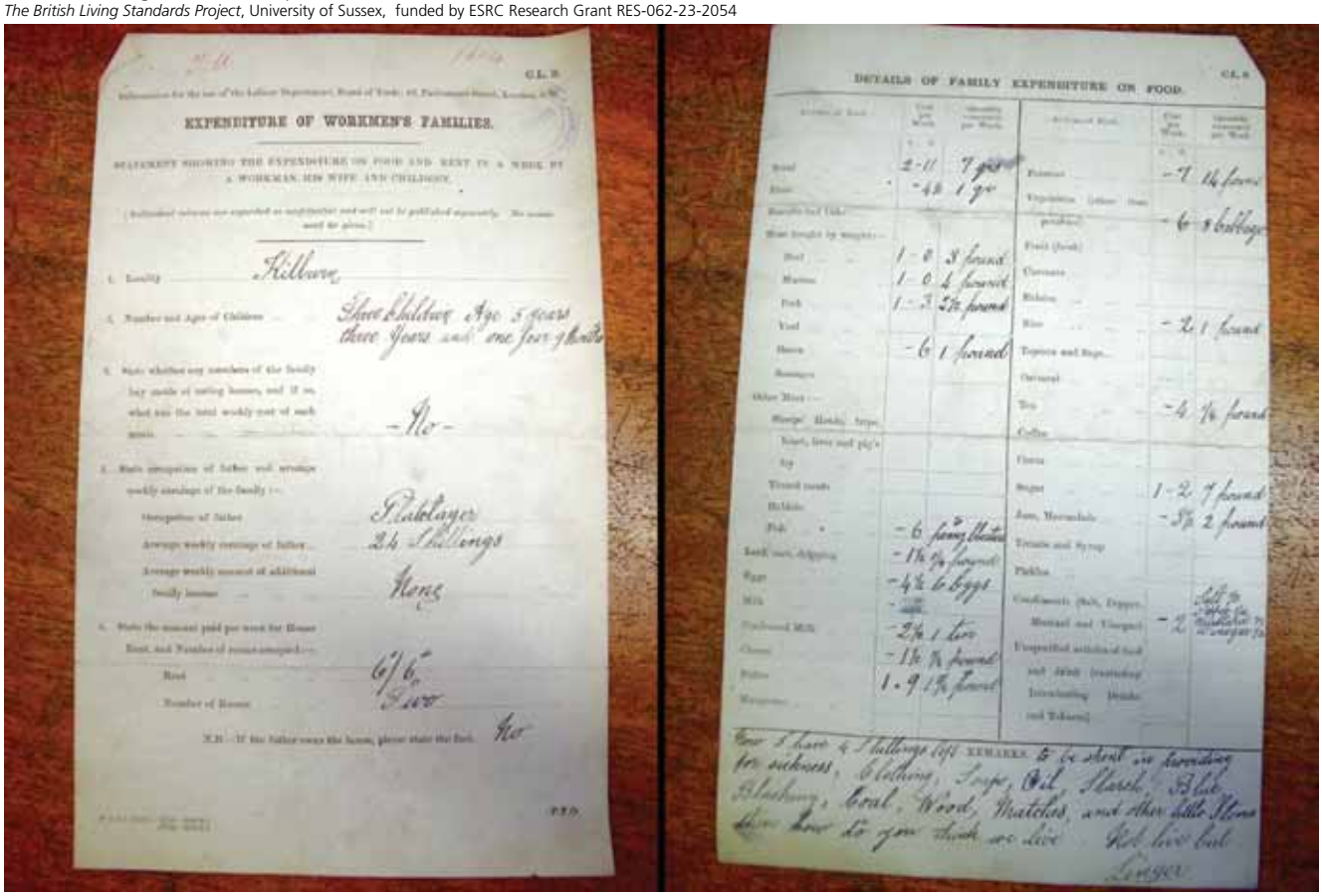
Similarly, Professor Mark Ormrod of the University of York and his colleagues have reconstructed the lives of numerous high-profile immigrants in England in the medieval period as a result of the digitisation of the medieval Alien Subsidy Returns and Letters of Denization in the England's Immigrants project.[2] Perhaps more interesting, they have used a prosopographical approach to collect the scraps of information in these sources about more humble individuals to develop a fascinating new insight into patterns and processes of migration in the medieval period. As Professor Ormrod himself observes:

> Electronic resources have transformed access to historical records for an increasingly diverse range of users, and thus made it possible, as never before, for people to understand the evidence on which we base – and extend – our knowledge of the past. The opportunity to reach public audiences has been especially important for medieval and early modern historians, because the records we use are written in hands and languages that are difficult to read. So electronic delivery of document facsimiles, modern English translations and summaries, and databases of standardised information have helped to revive interest and confidence in working with medieval records. Among some of the recent key examples here are the Fine Rolls of Henry III,[3] England's Immigrants, 1330-1550 and the York Cause Papers project.[4] Electronic access gives history a bright future!

Robot historians?

It may be that digitisation will continue to transform the discipline of history still further. Canadian historians Geoff Keelan and Kirk Goodlet have speculated a future in which historians will be aided and abetted by robots.[5] This does not mean a metal historian! It is entirely plausible that before long software will be able to read difficult handwriting and convert to more readable forms. Texts in modern and ancient foreign languages might be translated with sufficient accuracy. As more and more transactions are digital, so the ability of software to archive these transactions becomes feasible. Of course this may be a curse rather than a blessing as we are overwhelmed with data. Having said that, robot readers could conceivably 'understand' texts to the extent that they might be able to direct historians' energies by recommending them to study particular source materials. Artificial Intelligence may seem the stuff of science fiction but it is a rapidly-developing field and may open up new approaches to the study of history.

An example of a return for a railway platelayer in Kilburn, North London, in 1904. Records became more detailed in subsequent surveys. While individual records provide a fascinating insight into the lives of families, the digitisation of large surveys opens up the prospect of new levels of investigation and interpretation based on thousands of such records.
The British Living Standards Project, University of Sussex, funded by ESRC Research Grant RES-062-23-2054



What are the challenges in digitising a collection of sources from an archive and making them available?

The process of digitisation itself is well practised and understood, with four main steps to get documents from artefact to screen.

1. Preparation of material to ensure the best image possible can be obtained. This may require documents to undergo conservation work.
2. Photographing is usually done in high-resolution TIF format which ensures the digital image files contain as much data as possible.
3. Processing the image files into output formats to meet the requirements of the project. These will usually be either JPEG image files or PDF which can allow for optical character recognition (OCR) to aid with the transcription of documents if required. This stage also involves developing appropriate file structures and metadata which will help users to work with the material.
4. The final step is making material searchable to facilitate research. This can include any combination of tagging, transcribing, OCR and crowd-sourcing to make machine-searchable formats. This final stage often poses the biggest challenge as the nature of the documents can determine how it is undertaken.

'Structured data' such as forms are often easier to work once the key fields have been identified to aid with searching. Not everything needs to be made searchable – usually only key data fields such as name, address, date of birth, service unit etc, as with the First World War service records.

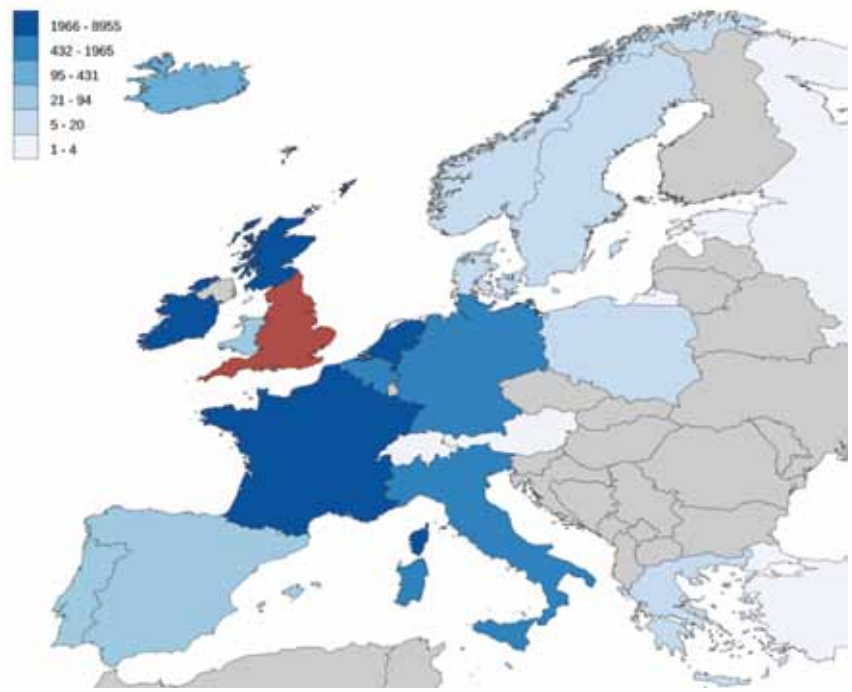
'Unstructured data' such as letters, minutes and reports are more challenging because what needs to be transcribed will depend upon the viewpoint of the user. This usually means transcribing everything. In the case of the Cabinet Papers project www.nationalarchives.gov.uk/cabinetpapers the typed minutes and memoranda could be transcribed using OCR and then a selection checked for quality control. The hand written Cabinet Secretary notebooks have to be transcribed by hand, however, which is time consuming and therefore expensive. One solution to this is to use crowd-sourcing projects such as Zooniverse who bring large numbers of digital volunteers to work on records by tagging terms such as names, locations and key events. The WO95 Operation War Diary has used this approach to create metadata about the official diaries kept by each battalion commander during the First World War.[6]

Bringing digital collections to wider audiences

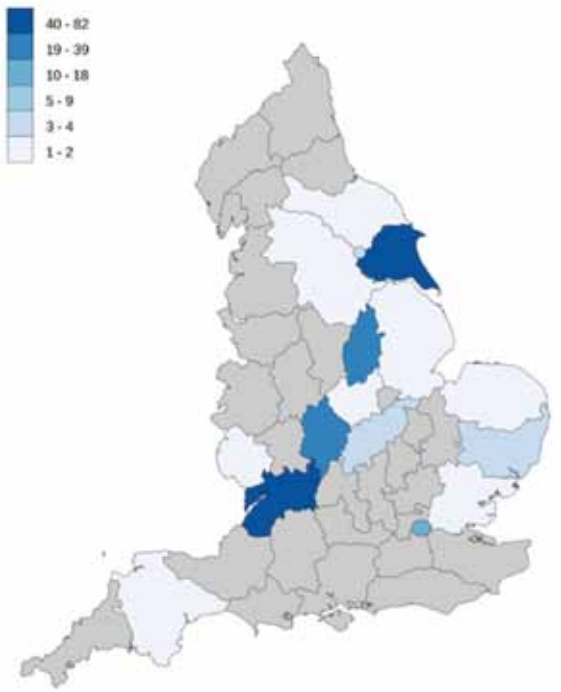
So far we have focused on what academic historians have gained, and might further gain, from the process of digitisation. But the National Archives and most other holders of collections do not want to reserve their collections purely for academic use. The last few decades have seen an explosion in public interest in history. Probably the largest single element of public activity has been in tracing family histories. In this respect the digitisation of large collections such as census returns and military records, particularly from the First World War, has opened up historical research to a completely new audience.

History teaching has also benefited from the opening up of collections in digital form. Primary and secondary schools make extensive use of original historical documents in their teaching of history. Some believe that school students cannot make proper use of such materials. We do not, alas, have space in this article to explain why we believe such naysayers are wrong, but they would be welcome to visit The National Archives, The British Library

Visualisations from the England's Immigrants project datasets. The graph on the left illustrates the points of origin of the majority of medieval immigrants to England for whom nationality is known. On the right is a representation of the number of Icelanders living in England in the medieval period.
www.englandsimmigrants.com/



England's Immigrants 1330-1550 (www.englandsimmigrants.com/, version 1.0, 29 January 2016



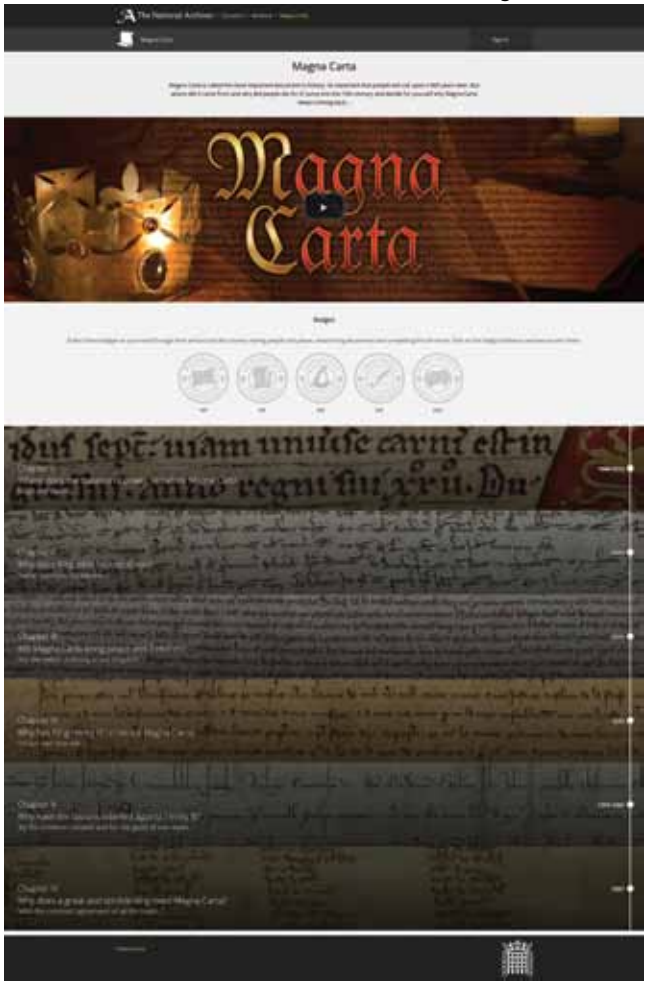
England's Immigrants 1330-1550 (www.englandsimmigrants.com/, version 1.0, 2 April 2015

and countless other local and national institutions to be proved wrong in person. To avoid any possibility of misunderstanding, we are not advocating a replacement of traditional teacher-led exposition or the demise of textbooks. We see the use of archival documents and other sources as one aspect of a good history education, complementing and enhancing the many other ways of teaching history which are available and which inform and enthuse students in the subject.

On the other hand, such materials as the Cabinet Papers or Magna Carta cannot simply be placed in front of young students. So a challenge which follows on from the digitisation of collections of documents is how to make these materials accessible to a general and/or a school audience. There are a number of key challenges. Perhaps the most formidable is language. Early documents are generally in Latin or perhaps medieval French. Even contemporary documents are couched in terms which can be inaccessible simply because of the formality of their language or because they are full of technical terms. Appearance is another potential barrier. The dense text of medieval scribes was once referred to as 'nothing more than scribble on a page' in one particular meeting relating to medieval collections at the National Archives!

Digitisation and technology is something which can help with this set of challenges. At the simplest level, technology makes it relatively easy to make transcripts (and or translations) available. Technology can also make different levels of transcript or translation available, with transcripts and simplified transcripts, for example. But technology can also help by making it relatively easy to provide additional supporting materials such as audio versions of documents. Technology can also provide digital equivalents to 'notepads' allowing students to record and share thoughts and ideas and questions easily in a medium with which they are comfortable and familiar. There are many examples where digital technologies have made apparently inaccessible materials accessible to a wide range of school students from all over the world. They range from the Norman Conquest to the Cold War and beyond.

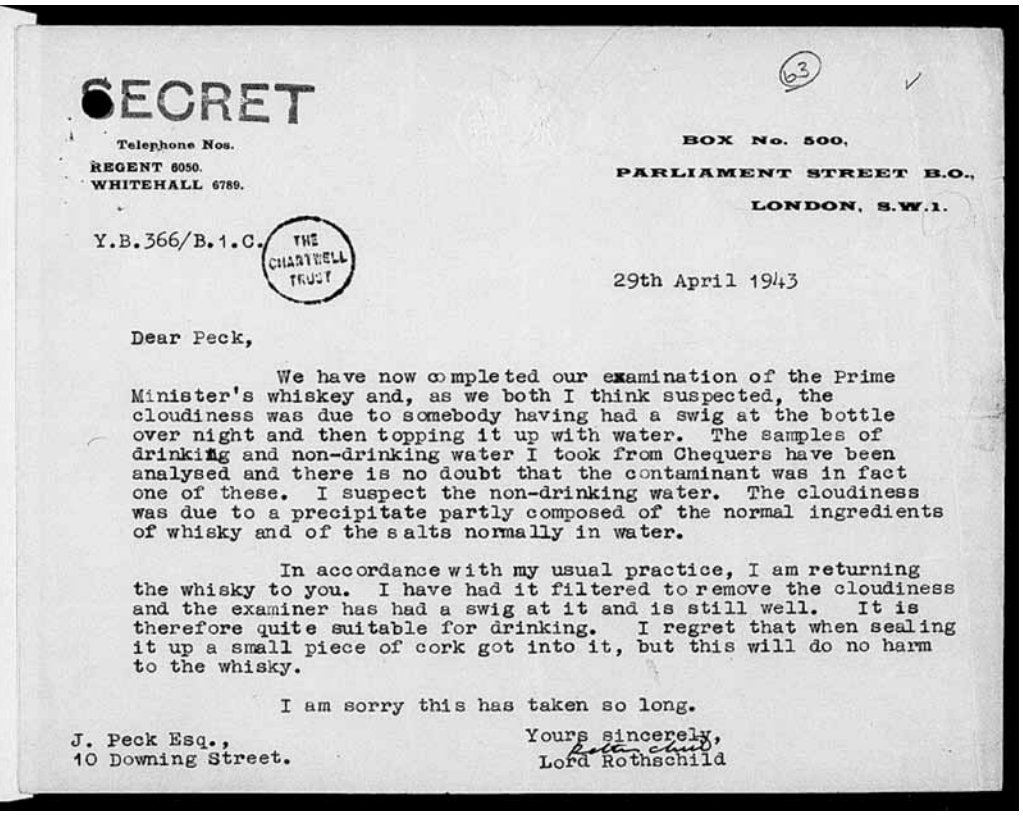
The National Archives Education resource on Magna Carta.



Magna Carta

In March 2015, The National Archives launched 'Magna Carta and the emergence of Parliament', a digital resource for school students developed in partnership with Parliament.[7] Part of the 2015 Parliament in the Making programme, it marks the 800th anniversary of the sealing of Magna Carta and the 750th anniversary of Simon de Montfort's parliament, and sets these events within the wider context of the struggle between kings and barons from 1066 through to the end of the thirteenth century.

A letter from the Churchill Archive concerning the contents of Churchill's whisky bottle.
By kind permission of Curtis Brown, the Churchill Archives Centre and Bloomsbury Publishing.
www.churchillarchive.com www.churchillarchiveforschools.com



SECRET

Telephone Nos.
REGENT 6050.
WHITEHALL 6789.

BOX No. 500,
PARLIAMENT STREET B.O.,
LONDON, S.W.1.

Y.B.366/B.1.C.

THE CHARTWELL TRUST

29th April 1943

Dear Peck,

We have now completed our examination of the Prime Minister's whiskey and, as we both I think suspected, the cloudiness was due to somebody having had a swig at the bottle over night and then topping it up with water. The samples of drinking and non-drinking water I took from Chequers have been analysed and there is no doubt that the contaminant was in fact one of these. I suspect the non-drinking water. The cloudiness was due to a precipitate partly composed of the normal ingredients of whisky and of the salts normally in water.

In accordance with my usual practice, I am returning the whisky to you. I have had it filtered to remove the cloudiness and the examiner has had a swig at it and is still well. It is therefore quite suitable for drinking. I regret that when sealing it up a small piece of cork got into it, but this will do no harm to the whisky.

I am sorry this has taken so long.

Yours sincerely,
Lord Rothschild

J. Peck Esq.,
10 Downing Street.

The resource requires students to work with over 30 original documents from the period to investigate how and why Magna Carta is issued, re-issued and evolves over time. Medieval documents present real challenges for engaging teachers and students, however, as they appear little more than illegible text.

The solution was to adopt an unconventional approach which requires students to work independently to write a comprehensive account of the struggle between kings and barons over a 250-year period. This requires them to read, understand, draw evidence from and substantiate judgements about four key points in time – 1215, 1225, 1265 and 1297 – before wrapping the whole thing up in the style of a medieval chronicle. The opening video presents the mystery of different versions of Magna Carta issued at different times by different people for different reasons and then set the enquiry: why does Magna Carta keep coming back?

Characterisation is with a cast of figures from the period to draw students into the period; to participate in a journey of discovery using ancient maps to track down parchment texts and press well-intentioned, if occasionally impertinent, questions upon powerful people. The great monk chronicler, Matthew Paris, acts as a guide and mentor; setting tasks, selecting documents, helping with translations and explaining their meaning but always recognising that ultimately the student is actually the master in the relationship. For while Paris, as with all of his contemporary chroniclers, is a dab hand at recording the events of the period, he is less capable in explaining *why* they occur and judging what their significance may be.

Students are rewarded throughout the resource with badges for visiting locations, reading documents, interviewing characters and completing chronicle chapters.

- Explorator (Explorer)
- Inquisitor (Researcher)
- Interrogator (Interviewer)
- Chronicator (Chronicler)

If they complete all this they will become Magister Chronicator (Master Chronicler) but more importantly they will have developed an incomparable knowledge and understanding of the complexity of how Magna Carta evolved and how it is linked to the rule of law, the genesis of our rights, the origins of Parliament and the foundation of our constitution.

The Churchill Archive

The Churchill Archive holds the personal archive of Winston Churchill.[8] By the standards of many archives it is quite small. But that still means around 800,000 private letters, speeches, telegrams, manuscripts, government transcripts and other key historical documents. These range from the iconic, such as the handwritten drafts and annotations of wartime speeches, to the mundane, such as the letters from Churchill's constituents on all manner of issues great and small. There is also the delightful, such as the exchange between officials concerning the Prime Minister's whisky. So the first challenge was to find a way to make this overwhelming amount of material navigable. In order to make this collection accessible to teachers and students in school Bloomsbury publishers commissioned a team of teachers to create investigations based on selected sources. The sources were presented in their original forms, something we believe is very important. An official document may appear intimidating but when it is stamped 'Top Secret' and the list of people to whom it was circulated includes the monarch, the Prime Minister and numerous other important people then it can generate intrigue. The sources were also supported by transcripts and simplified versions. Another challenge with the Churchill Archive was the perception that it could only be of value for studying Churchill himself or for great events. This was tackled through a series of short articles providing advice on how the collection could be harnessed for other histories, particularly by focusing on the letters written to Churchill by his constituents. These letters provided a treasure trove of insights into the preoccupations, perceptions and viewpoints of the ordinary people Churchill represented in his Parliamentary career.

REFERENCES

1 British Living Standards Project www.sussex.ac.uk/britishlivingstandards/protect
2 England's Immigrants 1330 – 1550 Resident Aliens in the Late Middle Ages www.englandsimmigrants.com/
3 Fine Rolls of Henry III www.finerollshenry3.org.uk
4 York Cause Papers www.hrionline.ac.uk/causepapers/
5 Robot historians and the future of history http://clioscurrent.com/blog/2015/6/29/robot-historians-and-the-future-of-history
6 WO95 Operation War Diary www.operationwardiary.org
7 National Archives Magna Carta www.nationalarchives.gov.uk/education/medieval/magna-carta/
8 Churchill Archive for Schools www.churchillarchiveforschools.com/

Ben Walsh is Associate Vice President of the Historical Association. He is a textbook author, trainer and senior examiner.

Andrew Payne is Head of Education and Outreach at The National Archives where he leads on the development and delivery of services for schools, teachers and communities.