The Challenge of Born-Digital Records Management at The UK National Archives

Mary Gledhill
Commercial and Digital Director, The National Archives, U.K.

Summary

Digital records management has created huge opportunities for archives, as well as posing some, as yet unsolved, major challenges. The UK National Archives has achieved significant progress in digitizing paper records for presentation online – approximately 300 records are now downloaded for every one delivered to the public in their reading rooms in west London. Their 'Discovery' service which launched in 2012 set new standards for searching records in archives across the UK and is at the heart of a service to users that includes remote record copying, e-commerce, tagging and many other features. The National Archives' website had more than 19 million visits last year, more than a third of them from mobile devices.

The National Archives has developed the infrastructure for permanently preserving digital records. They are now embarking on the difficult task of scaling up their processes in response to anticipated demand from government departments who will start transferring born-digital records in ever-larger quantities from 2016. Pilots with selected departments are underway which have highlighted some major issues for which no simple solutions are currently available. The volume of digital records, and the near impossibility of manually reviewing records in order to identify sensitivity requirements are the subject of their investigations. The approach of 'learning by doing' has however enabled the organisation to achieve early success in publishing the first open born-digital records on its website earlier this year.

This keynote will discuss what it takes to become a 'digital archive by design'.

Biography

Mary is responsible for the digital development which sits at the heart of The National Archives' online services. This includes our website which received more than 15 million visits last year, and our digital records infrastructure which allows us to collect and permanently preserve the ever-growing body of digital records produced by the UK government.

She also leads the commercial team which is responsible for business development and income generation, including major licensing and publishing partnerships, trading activities and the digitisation services that we provide to other organisations.

Mary has experience of working in a broad range of industries, starting as an engineer before moving into management consultancy with Booz Allen Hamilton following an MBA at Cranfield School of Management. She then joined BBC Worldwide, the commercial arm of the BBC, where she was responsible for strategy and business development activities for the Home Entertainment division. She joined The National Archives in 2011 as Commercial Director, adding digital development to her portfolio in 2015.

The Challenge of Born-Digital Records Management at The UK National Archives

Mary Gledhill

Commercial and Digital Director, The National Archives, U.K.

Good afternoon, everyone. My name is Mary Gledhill and I am Commercial and Digital Director of The National Archives.

Today I would like to talk to you about the impact of the digital era on The National Archives in the United Kingdom. I will discuss some of challenges that we face, together with the opportunities that we have identified and the innovations that we have introduced in response to the ever-changing digital landscape.

I will be very happy to take your questions at the end.

I would like to start by giving you a brief introduction to The UK National Archives.

We are the official archive and publisher for UK government, and for England and Wales.

Sponsorship of The National Archives has recently moved from the UK Ministry of Justice to the Department of Culture, Media and Sport. Our statutory responsibilities are derived from the Public Records Act, which mandates government departments to select and transfer records of historical value to The National Archives for permanent preservation. In practice the responsibilities under the Public Records Act are delegated to the 'Keeper of Public Records' who is the Chief Executive of The National Archives.

Somewhat unusually, we receive our funding directly from Her Majesty's Treasury. This position reflects the fact that we may, on occasion, be called to hold any department (including the Department of Culture, Media and Sport) to account for their record keeping under the Act.

The Public Records Act in the UK dates from 1958, and therefore makes no reference to the challenges of digital information management. We keep this apparent 'gap' under review, but to date we have found that the simple terms in which the Act is written provide a degree of flexibility that might be absent from more recently drafted legislation. We therefore have a mandate to collect, preserve and provide access to public records in any format.

One aspect of our operations under the Act has recently started to change. In 2013 the UK government began to move towards releasing records when they are 20 years old, instead of 30. Two years' worth of government records will be transferred to us and made available for public access each year until 2022. The entire transition to a 20-year-rule will take place over ten years.

The transition to the 20-year rule is a significant undertaking for departments, places of deposit and The National Archives. The fact that this transition is taking place at the same time as the transfer and permanent preservation of the first large quantities of digital material is the first of our major challenges!

The National Archives has around 600 staff, but we are a complex organisation as we also fulfil a number of other roles.

- We advise on information and records management in government and provide training for government bodies on cyber security. We also manage Crown copyright and publish UK legislation online.

- From a public perspective we aim to be a vibrant cultural and heritage institution, making our collections available for inspection and research in our reading rooms and online for customers all over the world.

- We have a mutually beneficial relationship with many UK universities and other research institutions, and provide award-winning educational sessions for school classes and training for history teachers.

- And finally, since 2012, we have played a leadership role for the archive sector in the UK, with responsibility for 2,500 institutions across the country. The sector includes both local public archives which have a role as ‘places of deposit’ for records of local historical value, as well as private archives.

Our approach to sector leadership is in itself quite diverse in its scope, and I will say more about the digital aspects later.

We are ambitious in our plans, particularly when it comes to increasing public engagement with the records that we hold, and seeking to ensure a vibrant future for the archive sector. However it has to be said that most public archives, including ourselves, have had funding cuts of between 25 and 40% over the last five years. The new UK government elected in May 2015 has indicated that similar spending reductions will be enforced over the next five years. This will inevitably place limitations on what can be delivered, even if we continue to succeed in making efficiency savings and generating income from alternative sources.

Over the last fifteen years, digital technology has changed fundamentally what it means to be an archive. We may have only relatively recently started to accession born-digital records, but we have digitised and published online around 144 million historical documents. This equates to between 8 and 10% of our current paper collection.

Digitisation is expensive – prohibitively so for rarely-used documents – but for those documents most in demand it creates a new opportunity to provide global, rather than local access to records. Current statistics show that around 300 records are downloaded from our, or our partners’ websites, for every one paper record that is accessed in our reading rooms. Thanks to digitization, research can be carried out anywhere in the world, at home or on the move.

For financial reasons, our relationships with commercial online publishing partners have been critical to the success of our digitization programme. The investment from partners in digitizing records over the last fifteen years has far exceeded the amount that we would have been able to fund ourselves. In addition, by returning royalties to The National Archives for the records that they license, our partners have helped us to fund our core in-house services in the face of cuts elsewhere.

Our partners can bring deeper knowledge of specific audiences, such as genealogists or academics (whereas we, on the whole, aim to be generalists). They invest time, money and expertise in acquiring and serving their customers to an extent that we could not hope to replicate. As such, while we have our own website and download services, commercial partnerships are a well-established way for us to broaden access to our records online.

One final point – when we first started to undertake bulk digitization projects, there was an idea that over time this would reduce the demand for access to original documents in our reading rooms. This has definitely not proved to be the case. What has happened instead is that users do the early parts of their research online, using digitized material, but then almost inevitably reach a point where the next step in their search is only available on paper. Our visitor and production numbers are as high as they have ever been, but the record series being viewed have changed.

The other opportunity to transform the experience of archive users has been the development of our online catalogue. This has evolved beyond recognition over the years. The last major development at The UK National Archives was the launch of the multi-faceted service that we call 'Discovery' in October 2012.

The launch and subsequent upgrades of Discovery have produced a service which

- Brings various databases of archival reference information together into a single entity where users can search and browse over 30 million record descriptions.

- Uses approaches which are familiar to non-expert users, such as a single search box and filters to narrow down their searches.

- Also features advanced search functionality and guided searches of particular record types for more expert / specialized users.

- Is based on Lucene open source software.

- Allows users to tag records making them easier to find again.

- Incorporates 10 million record descriptions from 2,500 other archives around the UK, together with details of those archives and their collections. Previously this information was held in separate databases.

Provides a seamless user journey, from finding a record in our catalogue, to the various access options available. Depending on the record in question, the options include pre-ordering the paper record for viewing in our reading rooms, requesting a copy of the record to be produced and emailed to the user, or in the case of 9m records that have been digitized in bulk, paying for and downloading a digital copy. This is a significant improvement on the previous arrangements where the online catalogue and download systems were separate and therefore required a lot of 're-keying' by users wanting download records that they had found in the catalogue. As discussed previously we also link from Discovery to partner websites where other digitised records from The National Archives can be viewed and downloaded.

The evolution of Discovery provides an interesting example of the impact of digital technology from an organizational perspective, as well as a purely technical one.

Prior to the development of Discovery, our digital systems at The National Archives had evolved from a variety of different disconnected initiatives, in a way that is not untypical of organisations going through the early stages of digital transformation.

So the cataloguing team continued to 'own' and make decisions about catalogue content as they had when the catalogue was on paper. Choices about which records should be digitized and made available online were mostly made by our e-commerce team – apart from the odd occasion when alternative funds were made available, in which case the e-commerce team would still be expected to provide remote customer support. On-site customer support was provided by a different team in the reading rooms. Last but not least, the technology team, as they were the ones building the website, tended to control the functionality delivered – sometimes, but not always, aligning their priorities with everyone else!

These early initiatives were invaluable for our digital development but left us with a confusing user experience, and lots of opportunities for decisions made by one team to have an adverse effect on the activities of another. The decision to bring together the catalogue with the e-commerce service and other access options helped to dramatically simplify the user experience. However it also magnified the potential for unintended consequences as a result of fragmented decision-making.

In my experience this is quite common when businesses try to adapt to digital ways of working. The organizational boundaries which worked well in a physical world break down in the digital world, and either greater collaboration or a whole new business structure is needed in order to create efficient and effective digital business processes.

At The National Archives, we addressed this issue first by creating a new steering board for 'Discovery' which involves all the relevant parties in decision making. The insights from that board have also influenced our subsequent decisions about the wider structure of the organisation, as we seek to become 'digital by design'.

Discovery is now a very powerful tool, although the first few months after the launch of the new service were challenging as we received a lot of criticism from users of the old catalogue. Many of them felt that useful features from the old system had been 'lost' or were harder to use in the new service. In a few cases we have now acknowledged the issues raised and built new tools in Discovery to replicate specific functionality.

However we have explicitly *not* tried to recreate every feature of the old system. While some of our most expert archive users found it difficult to adapt at first, we have very deliberately focused on making Discovery easy to use for all – and that has meant simplifying the core product. More specialist searches can still be carried out by applying different tools, such as filters or advanced searches.

Discovery passed the UK Government 'Digital by Default' service standard assessment with very positive feedback earlier this year. Our web analytics show that more users, including many people who have not used the website before, are finding relevant results. However as we now update Discovery on a regular basis, helping users to understand site changes and get the best out of their searches is a continuing challenge that extends across our web editorial, customer service and communications teams.

The next phase in the development of Discovery will focus on three main areas:

i) Making the service (and the rest of our website) fully responsive. Almost a third of the 19 million visits to our website last year were from portable devices.

ii) Enabling other archives to manage their own catalogue references on Discovery remotely. We are developing a suite of tools for downloading and uploading references both individually and in bulk, to enable archivists to add new records and make corrections where required.

iii) Providing access to born-digital records. The first born-digital records were made available on Discovery in August this year.

It is this third point that I will cover in more detail during the rest of my presentation.

The first iterations of digital preservation systems at the UK National Archives were built as early as 2003 to preserve relatively low volumes of widely varying digital files using our in-house research and experience. At the time, contemporary thinking about digital preservation held that significant human involvement would remain an essential part of the process of accessioning and preserving records.

By 2011 it was apparent that the digital landscape looked very different. Born digital information was being produced in vast quantities, both in government and elsewhere. The challenge was now to find ways of dealing with large numbers of digital files, mostly of similar type, using processes that require minimal human intervention.

Our response was to start building a digital records infrastructure or 'DRI system' that could handle the long term storage of both digitized and born-digital records. We have successfully been ingesting digitized material into DRI since 2012 and now have terabytes of digitized images in the system. Along the way we have learned valuable lessons about how to manage and move around large volumes of digital information, how to automate some of our QA processes and how to limit the risks posed by viruses and other security threats.

We began to ingest a few born-digital records in 2013, but each addition still needed bespoke arrangements and many manual interventions. It was last year, in 2014, that we first began to really engage with the challenge of what it would take to accession born-digital records 'at scale'.

The work has been led by The National Archives, but has gained additional attention thanks to the commissioning of Sir Alex Allan, a former senior civil servant in the UK, to conduct an independent review of information management across government in 2014, followed by a further review of digital records management this year. His recommendations will contribute to the evolution of plans for the future of records management.

The first task for the team considering our born-digital records transfer process was to determine the scale and features of the digital records landscape. This was more difficult than it may sound. Looking back to the early introduction of digital systems in government, more than 25 years ago, it was apparent that different departments introduced information management systems in different ways and at different points in time.

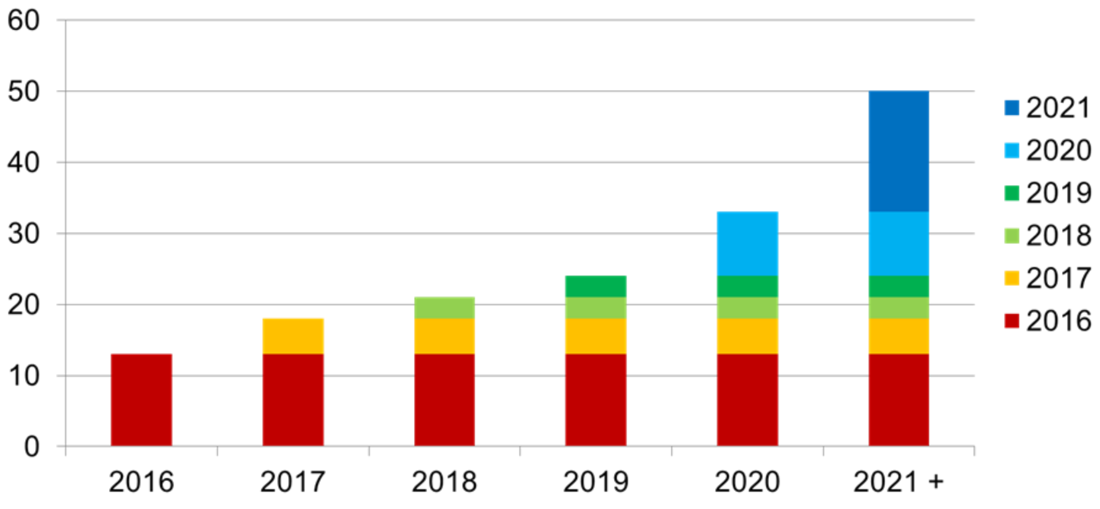
A mixed picture emerged:

- At a time when digital networking and storage were both relatively new, the principles of managing information assets and migrating them from old to new systems were less well understood and followed than they are today.

- Some Departmental Records Officers, who had clear responsibility for the filing, retention and eventual archiving of paper records, found it hard to get involved in electronic records management which was often under the control of IT teams.

- Perhaps as a result, some departments initially implemented 'print to paper' strategies – producing 'official' paper copies of documents to file and keep while still, in many cases, retaining the digital versions.

We therefore expect the next 5 years or so to be characterized by the transfer of increasing quantities of digital records, but with a limited decrease in the quantity of paper records transferring on an annual basis. We are calling this the era of 'hybrid' transfers. The specific challenges of managing hybrid transfers are apparent if you consider that the 'same' record is quite likely to be captured in both physical and digital form, but with different metadata.



The graph shows the number of departments that expect to make digital transfers to The UK National Archives in each year. You can see that the curve rises steeply – this effect is even more pronounced if the quantity of digital records to be transferred is also taken into account.

Based on the intelligence collected from government departments about their digital records readiness, we decided that our work for the next few years should focus less on the development of new technology and more on the challenges of scaling up the *processes* associated with digital transfer.

We realized that we needed to dramatically reduce the amount of manual input to the transfer process, both from The National Archives and from the government departments who are responsible for selecting and sensitivity reviewing records for transfer.

Given the major uncertainties around born-digital records, we agreed that we would approach these problems by carrying out a group of representative 'pilot' projects to find out what issues actually arose in practice, as opposed to what issues were considered significant in theory.

We initiated six pilot projects. Each was selected to highlight a particular set of issues or challenges. These included:

- records written in English and a second language (from the Welsh Government)
- records extracted from an EDRM system (our own objective records)
- potentially sensitive video material from a recent inquiry (Al Sweady)
- hybrid records from the Foreign and Commonwealth Office raising particular challenges around contextual sensitivity review
- a dataset from Companies House and
- a significant quantity of video material (much of it very long and not very interesting) from the UK Supreme Court

The sets of records involved were chosen to be of a sufficient scale to be realistic but not so large that they would totally overwhelm our team if problems were encountered. Not surprisingly, we did run into lots of issues – however this did not discourage us as, in solving problems for the pilot record sets, we have produced solutions that could be applied more widely in future.

The first two pilot record sets have successfully gone through the entire end to end process, while work on the others is continuing. The National Archives' technical capability to ingest new content is now relatively well proven, so some of the most significant challenges are around selection and sensitivity reviews, which are carried out by the contributing department.

Our research indicates that more than half of the top 21 government departments have not yet even considered how to carry out digital sensitivity reviews. The process of sensitivity reviewing paper records is already rather time consuming, so applying traditional approaches to the huge volume of digital records produced by government is daunting to say the least.

However without viable and effective sensitivity reviews, our ability to open records to public access will be severely limited. This would undermine the whole purpose of the Public Records Act and The National Archives so it is an area where we must take the lead in finding solutions which can be used across government.

The most common type of sensitivity by a significant margin is personal sensitivity. This is even more skewed if you take a department by department approach, as most records containing information of national interest are concentrated in a relatively small number of departments such as the Foreign and Commonwealth Office.

This is, to some extent at least, is good news, as our research suggests that personal data may be relatively easy to identify using commercial software solutions. These can be applied to records at scale and should identify certain types of sensitive information automatically. We are in the process of trialling a number of different software solutions to evaluate their effectiveness in finding personal data.

Sensitivity based upon context is much more difficult to identify. This would include, for example, the sort of information that may be contained in FCO records and other material of national interest. We would be particularly interested to talk to any other national archives that are considering this issue as our investigations to date have not revealed any meaningful alternative to the use of expert reviewers looking at one record after another. Given the volume of digital records being produced and the pressure on government to cut costs, it is very difficult to see how such a manual process will be even remotely sustainable in future.

Finally in this section it is worth mentioning that we also collect and preserve the UK Government Web Archive as we consider that, in a digital age, this also represents an important channel via which government departments communicate with the public.

We have been archiving websites since 2003, but the number of sites and the frequency with which they are crawled was greatly expanded in 2008, in order to support our own Web Continuity project aims, and the then-government's website review programme. This programme involved improving efficiency and reducing expenditure on online services by closing government websites and bringing information together into ‘supersites’. We were tasked with capturing all closing websites, including those featuring public inquiries.

Another wave of website closures has taken place since 2011, as the current Government Digital Service led a transition to the new single government domain GOV.UK. Our Web Continuity team worked again with government departments and collaborated with the Government Digital Service to make sure that older government content remained accessible. Websites were preserved each month, enabling users to be redirected to the UK Government Web Archive wherever content had not been migrated onto GOV.UK.

The UK Government Web Archive is used by more than 1 million people every month. Last year, we collected 336 million URLs and preserved an additional 10.3 terabytes of data. We work closely with The Internet Memory Foundation, who undertake the crawling, hosting and much of the technical development of the service under contract to The National Archives.

We are currently embarking on a new piece of ‘user needs’ research for the web archive in order to identify the likely usage patterns of the future and inform decisions about the development of new tools and features. We are particularly keen to explore further integration of the web archive with Discovery.

Without doubt, the archiving of digital records is a major challenge for The National Archives. Our responsibilities towards our collection of paper records will not diminish in any way, and yet the additional workload associated with digital records will be significant. Finding a way to manage this with decreasing funding from government is bound to be difficult.

There may however be some opportunities on the horizon too. The UK Government Digital Service is advocating a ‘cloud first’ approach to procuring new systems, which may help us to increase the scalability and the security of our digital services.

We are gradually adding to our portfolio of solutions for the handling and presentation of digital records. Our approach to presenting born-digital records in Discovery, for example, appears to be working well. We have decided to create a new reference for each record but include as metadata the original file names and file structure that surrounded it. There were a number of examples from the pilot where the metadata provided essential context to support understanding of the record itself.

We will continue to keep under review the wider questions around record transfers and the Public Records Act. For example, does it make sense for digital records to be held for twenty years by the originating department before they transfer to The National Archives? What date should trigger the start of that twenty year period? And what is the role of local ‘Places of Deposit’ in providing storage for, and access to, entirely digital collections of records?

Although records management in the UK has some unique aspects, we know from our peer review work that we share our overall digital challenges and opportunities with archives around the world, as well as with smaller archives across the UK.

We are well aware that the issues we face are probably too big and too difficult for us to solve alone. Collaboration and partnerships are therefore a major part of our future strategy, because we know that we are likely to achieve more by sharing our best work with others.

I would like to thank you for the opportunity to discuss the progress made so far by The UK National Archives in the world of digital archiving.