Mass digitization of historical records for access and preservation

For every document that we deliver to readers of The National Archives at Kew, over 150 are delivered via the web. Currently there are over 60 million individual documents digitized, categorized and in many cases transcribed and available to download digitally. Capturing the business of Government has also changed – we have started archiving the e-mails, the Word documents and the websites of Government instead of the memoranda and bits of paper of the past – seamlessly ingesting these into our digital archive, and making that available through our website. The task of digitizing the collections that reside on over 175km of shelving is not an easy one, however, nor is it something to which there is necessarily a right – or a wrong – approach.

DAN JONES
Head of Business Development
The National Archives, UK

Academic researchers may associate The National Archives with studious historical research, the examination of primary source documents and, of course, with being the home of the *Domesday Book*. Other groups may associate the Archive more with family history research, because we hold the censuses and all the other sort of records you need to research your ancestry.

The reality however is that, increasingly, we are a digital archive and if you pursue family history online you could easily be using our content without ever visiting our site. For every document that we deliver to readers in the reading rooms at Kew, over 150 are delivered via the web. Currently there are over 60 million individual documents digitized, categorized and in many cases transcribed and available to download digitally.

Capturing the business of Government has also changed – we have started archiving the e-mails, the Word documents and the websites of Government instead of the memoranda and bits of paper of the past – seamlessly ingesting these into our digital archive, and making that available through our website. The task of digitizing collections that reside on over 175km of shelving is not an easy one, however, nor, as I describe below, is it something to which there is necessarily a right – or a wrong – approach.

Why should I digitize?

This may seem a question with an obvious answer. However, I think it is easy for cultural institutions like archives sometimes to get sucked into digitization, and that they feel pressure to digitize their collections without perhaps pausing to think what is motivating them, and then using that information to shape their approach. By that I mean how does digitizing their holdings fit with their vision, their business model, and what specifically do they need to get out of digitization initiatives? In the case of The National Archives the desired outcomes are clear:

- to create high-quality digital surrogates to ensure preservation of original documents
- to deliver digital documents over the Internet to maximize access
- to use digital technology to add value and aid interpretation
- to optimize online and onsite services along lines of agreed stakeholder segmentation.

If one puts all these points together, it becomes clear that only wholesale digitization of complete collections really delivers all these benefits to the archive. If we consider the classic academic publishing model of a few years ago, selecting documents and building up a microfilm or digital resource around a theme – that may increase understanding, it may deliver progress towards increasing access, but it does not deliver those other operational benefits. We cannot put those original documents in long-term storage; we cannot keep them away from public inspection because the digital resource is not a complete surrogate.

External drivers

On the other side of the coin are the external drivers affecting our approach to this area. 'itunes', 'Sky+' and, more recently, the BBC 'iplayer' have all shown how well technology can work, and this raises expectations for everyone else. Google is the prime example: so universal and all-pervading is familiarity with its search technology, it has become the benchmark for any search application anywhere. People used to be happy to look at paper catalogues, which would refer them to a supplementary finding aid, which would eventually take them to a box in our reading rooms. Now people expect to search the text of a document, and they expect contextual relevant results to be categorized and delivered in a way that makes sense to them.

It may have become a somewhat hackneyed phrase, but 'Web 2.0' and social networking also look like being a very creative force – for an organization like The National Archives which is founded essentially on the expertise of its staff, the idea that people are as willing to accept the wisdom of crowds as they are the wisdom of experts is a radical – and challenging – concept for us. More on that later…

It is well documented that the financial position in the public sector is quite tight, and I think it will be for the foreseeable future. A rough estimate is that it would cost about £5 billion to digitize our entire archive, which represents a big reality gap in terms of what people expect, and what we are actually able to deliver. This just emphasizes the importance of successful partnership and co-operation across both the public and private sectors if we are going to make significant progress.

It can get slightly disheartening to focus too much on the scale of what we are trying to achieve relative to what we have actually done to date. We may have over 100 million document images online by 2011, however, that will constitute only about 8% of the archives holdings. It reminds me of the story about the great French marshal, Hubert Lyautey, who commanded his gardener to plant a tree in his garden. When the gardener complained the tree would not reach maturity for 100 years, the marshal responded: "In that case there is no time to lose, plant it this afternoon". I think it is really important that, as we embark on individual digitization projects, we have a vision as to where this is leading in the long term, what's being achieved by each individual project and what our delivery model is going to look like in the future.

Models of digitization

For the models of digitization we use, broadly speaking, a mixed economy. We have our own delivery mechanism, something called Documents Online, which we use to deliver both internally-funded projects and projects for which we have sought outside funding. Then we have an extensive network of commercial partnerships, usually major projects that if we delivered them ourselves would have a huge effect on our infrastructure, or where the investment required is beyond The National Archives alone.

That many services will be paid for is a vitally important point and one that I know is probably quite controversial for some people, but I think if you segment and understand your stakeholders and your user base, it does emerge which services can be charged for, which can be free at the point of use to the individual but charged for along the line – by the way of a library subscription, for example. It also helps you understand which services, if they are to serve their intended purpose, really do need to be completely free. By doing this, it almost self-selects the appropriate method of digitization in each instance. I am also a big believer that the right business model is not commercial partnership, or internal funding, or grant funding or using the third sector or any other model in isolation. It is all of them, and the relationship between them is crucial if you are going to be able to digitize as much material as possible.

Terms and conditions

All contracts granted should be non-exclusive. What we are looking at here are ways of getting over the primary barrier, i.e. that the material is in paper form. If we can get that into the digital space avoiding any exclusive or preferential rights, that main barrier is removed and there is nothing to stop other parties building an equally innovative service to serve different stakeholders. The example here would be the census. Our primary aim may be to deliver a consumer service to allow family historians to use the material but once those digital assets are created, creative things can be done with those censuses in the academic space.

A good example among our current initiatives is the 1911 Census. It is a huge project: we have five automated scanners running pretty much round the clock, there are 18 million individual document images in total, which is about 40,000 images per day. Once the images are scanned, they get beamed to the Philippines where they are transcribed and five days later they come back fully quality assured. When it has been completed it will be approximately half a petabyte of data, so in virtually every way it is a massive project.

It is also a commercially attractive project. We had a lot of interest from a number of areas and this enabled us to secure additional services in the main contract. So once the consumer service has rolled out, which should sate the appetites of the country's hobbyist family historians, other services will roll out to academics and there will be specific access for schools. Postcode mapping will take place for the content, and statistical analysis search screen will be available as well. From my perspective it is an example that partnership with the commercial sector can deliver huge benefits right across the board to all our stakeholders.

Organizational impact

Harder to do from an organizational perspective is to trust third parties with uniquely valuable material; in many ways it goes against the grain – against the culture of an organization like The National Archives. You have to put a huge amount of resource into approving equipment, processes, managing those processes, training the staff, and that requires a deal of co-operation from professionals across the organization. And, fairly obviously, it is sometimes tricky balancing all these interests and making sure that you enjoy as much of the upside as you can.

The organizational impact is really an attitudinal change as much an anything. As an archive we are very used to providing; providing records, providing expertise, and now we are really moving to a model where we are enabling others to provide because they are better placed. That means the existing providers of resources, whether that is a subject specialist in our records knowledge department, or conservators in our collection care department, have to change what they do quite considerably. Luckily, colleagues at The National Archives have been very quick to recognize that the best form of preservation is digitization, and have supported this change of emphasis.

The final point is really to remind any institution looking to do something like this that the idea that partnering with third parties means you do not need to expend the resources yourself is a bit of a misconception, I'm afraid. Partnership means that you would still need large amounts of resource behind this; you would just need to employ that resource in a different way. Business development, management finance and strategy resources typically do not reside within archives, but they are required to effectively negotiate and manage contracts that deliver long-term benefits.

Joining the dots…

An undoubted side effect of this multi-dimensional approach to digitization is that it can give rise to a fragmented user journey for the committed user of National Archives content, with different material requiring visits – and transactions – with several different commercial websites. This comes back to my point about the main barrier being conversion from paper to digital files, and the application of metadata to enable search – the user journey – is a secondary issue, for The National Archives at least. It is almost a 'nice to have' problem, a by-product of our successful conversion of huge amounts of material. That is not to say we cannot begin to address this too, however, and there are a couple of developments which may help that we still believe we are best placed to provide ourselves.

The first is search. I think whilst there are some amazing things that Google can do with their search engine, the issue with our collection is really that the more results you return, the more baffling it can be for the user. So using Autonomy search technology we are developing our own kind of 'global' search for the website. This uses the expertise that we have within the organization to try and categorize and make sense of the results that it pulls back. It also means it searches all the website, databases and aspects of our electronic resources and presents them in an intuitive and logical manner. We think this is a key development and hopefully over the next few years will help us make sense of all the rich digital assets that we have managed to create.

Something else we are also very proud of is our 'wiki'- based resource 'Your Archives'. Not only have we been pleased with the keenness of our users to actually share the research and the expertise that they have spent years learning to others, we have also found that our staff are using it – effectively putting down what is in their heads. Something initially intended for our users is now enabling what is very much a 'one to one' kind of expertise model in the reading rooms to become a 'one to many' model via Your Archives. I think this is perhaps the most exciting bit about it, that like a lot of good digital developments, it is not just about putting more stuff on the web; it is really about re-engineering the way that we deliver our services. At the moment it is pretty much a stand alone application but I am quite sure that virtually every new resource that we develop going forward will have a link to Your Archives and that, increasingly, it will become a hub for people to interpret and make sense of our resources.

Results

So in terms of the results, what has this programme delivered for The National Archives? If we consider the number of National Archives documents delivered over the web from a zero base in 2002, last year there were over 80 million and we are forecasting nearer 100 million this year. By 2012 we will have 100 million documents transcribed and available to download. If we had looked to develop these services ourselves, it would have cost us around £40million and certainly taken a lot longer than five years, so it has been a very positive programme all round.

And finally let's just not lose sight of why we do all this. It is really to maximize access to an understanding of the incredibly rich material that we hold. So I know there are some compromises with the delivery model, there are some compromises with the services that we get at the end, but in terms of overcoming that huge barrier of having millions of pieces of paper and making sense of those in a digital space, I think we have made a strong start.

■ **Dan Jones**
Head of Business Development
The National Archives
Kew
Richmond
Surrey TW9 4DU, UK
Tel: +44 (0)20 8392 5206
Fax: +44 (0)20 8487 1974
E-mail:daniel.jones@nationalarchives.gov.uk
Website: www.nationalarchives.gov.uk