Digital information and the digital document[1]

David Thomas (The National Archives)

The story of mass digitisation of humanities resources probably begins in 1998 with the establishment of the New Opportunities Fund under the National Lottery Act. The government set out policy directions for the Fund in the same year, including one with a curious title: *Information that supports lifelong learning into digitised form*. The vision was that "through the digitization awards by the fund, people in every walk of life in communities across the UK will be able to connect to almost limitless gigabytes of resources ranging from the treasure-store collections of leading museums, galleries and libraries to priceless archives of film and the arts. The funding will help give people access to the full tapestry of UK society past and present……and even allowing people to take virtual walks through local and national heritage locations". One of the leaders of the New Opportunities Digitisation project described it in Chairman Mao's words as *letting a hundred flowers bloom*. So – carried on a tidal wave of Maoist rhetoric and New Labour money, humanities digitisation was born.

This radical vision was implemented very quickly and by 2003 there were 155 projects, at a cost to the Lottery Fund of £50 million. Although a number of these were not digitization projects, most were and there are some major successes amongst them, including Old Bailey Online, Humphrey Southall's Vision of Britain which contains a great store of travellers' accounts of their journeys round the country, British Film Institute Online and the National Archives and partners' Moving Here which is a valuable source of information about the experience of immigrants in the UK.

In 2001 there was another significant development with the launch of The American Family Immigration History Center, which makes the 25 million immigrant arrival records in the Ellis Island Archives available online. At pretty much the same time, the National Archives in the UK launched its digital version of the 1901 Census, although every schoolchild knows that the Census crashed when launched and did not stagger to its feet for another 10 months, it did stimulate the market for delivering family history online. When we tendered for the contract we were able to shortlist four bidders of whom three withdrew – two because they considered the rate of investment return was

[1] A lecture given at the Digital Connections workshop at the IHR, 31 March 2011.

not acceptable. However, once Ellis Island and the Census had proved themselves, other players began to enter the market for internet-based historical censuses and other family history resources and we now have two organisations selling copies of the 1911 census online.

One of the curious features of humanities digitisation in the UK is the way in which it has been funded – there are at least four separate funding streams. First, many institutions do some digitisation using their own resources; TNA, for example has digitised its collection of wills using its own funds. Some records digitised in this way are provided free, others for a modest charge for downloading individual items. Second come the funding bodies – in the UK these were originally the lottery – but this has gone away to be replaced by the Research Councils and JISC – to whom we are deeply grateful. Third are the academic publishing companies – Gale, Adam Matthew, Proquest, etc who often digitise specialised collections and rely on selling them in small volumes but at high prices. Fourth come the family history publishing companies – Origins, Ancestry, Find My Past etc who digitise records of interest to genealogists and whose business model relies on a mixture of subscriptions and selling high volumes of individual documents at low prices. The situation is even more complex in the world of book digitisation since we have the work of Google, the Internet Archive and the wonderful Project Gutenberg which relies heavily on volunteers and generates revenue from voluntary donations.

If digital is a revolution, then it is most like the Spanish Civil War where, on the Republican side there were a range of groups with quite different and potentially competing agendas.

But will the revolution continue to prosper or will it end up like 1660 in England? The existence of so many models for digitisation leads to some questions about the sustainability of digitised collections. Both grant-funded and commercially-funded resources are at risk. Large funding bodies insist that grantees have to guarantee that their material will be available for, typically, 7 years. We know that of the 155 projects funded by the New Opportunities Fund, 25 can no longer be found, while there have been no changes or enhancements to a further 83. Of the 155, there are only 30 which have been enhanced or added to since the launch.

There are two significant threats to the survival of digital collections – financial and technological. Guarantees of long-term availability of materials given with great enthusiasm and genuine commitment when the grant application is being completed can ring a bit hollow seven years down the line when the money has long since run out. The National Archives has recently been approached by one large project which was funded by a grant. The funding has now been used up and the team are dispersing. The organisers would like TNA to take responsibility for the resources which they have created. Equally, commercial organisations cannot guarantee to be around forever. Proquest originally known as University Microfilms has been around since the 1930s but other microfilm houses have disappeared over the years.

The problem of what to do when the grant money runs out is compounded by technological issues. While the National Archives would like to help ensure the survival of digitised material, we do not necessarily have the funds to acquire and preserve large collections of digital images of TNA records particularly when they have been generated using proprietary software. Nor do we necessarily have the skills to ensure the survival of such systems or to hand-craft the transfer of a digital resource from a university to Kew.

There has been some interesting work by the Strategic Content Alliance on business models to ensure the survival of digital collections. A 2009 report produced the Alliance found *that projects are experimenting with and have deployed a wide range of revenue generating models while at the same time finding ways to minimize their direct outlays by reducing the scope of their work or by taking advantage of opportunities for assistance and subsidy from host institutions and outside partners. So, at this stage of their development, most of the projects covered in this collection of case studies rely on a mix of generated revenue and host support. While a couple of them have been around long enough to demonstrate financial viability, for most of the cases we studied it is too early to tell whether the mix of sustainability strategies employed will succeed over the long run.*

As well as exploring business models for sustaining digital resources, the funding bodies need to consider technological issues. It seems likely that in some cases digital resources will have to be transferred to other institutions than the creating ones to ensure their survival. If that is to happen then funding bodies need insist that projects

meet minimum standards for resource creation using open-source software with the aim of ensuring that digitised records can easily be transferred to a new holder in the future. And if they claim that they are already doing this then their message is not getting through. Even this year we at TNA saw a copy of a grant application where the bidders were proposing to meet the sustainability requirement by simply sending us a database on a CD. The problem is not a theoretical computing issue – it is a live problem now. A friend of mine recently wrote a book on child poverty and was surprised to discover that between drafting her manuscript and the publisher getting to work on it, resources she had cited had mysteriously vanished.

As well as questioning whether the digital revolution can be sustained, we need to ask more fundamental questions about whether digitisation democratises access? There are two issues here. First, how is the material to be digitised selected? Some people claim that selection is under the control of large-scale commercial organisations who simply digitise the obvious resources. The New Opportunities Fund approach was that it should digitise a wide range of resources from a lot of institutions. In fact both positions have a degree of validity. Yes there are a small number of large digitisation companies who have digitised the obvious family history and other resources. On the other hand, funding programmes such as the New Opportunities Fund and the AHRC project have allowed a large number of flowers to bloom.

The second question is whether digitisation has allowed non-academic users to be able to use a range of resources for their own research pursuits. Clearly the provision of online access to family history records for a modest charge has greatly enhanced the use of them made by family historians, many of whom begin their research by using the big family history databases such as Ancestry. On the other hand and probably inevitably, the commercially funded academic projects are only of benefit to those who have access to university or national libraries.

The picture for digitisation projects funded by the major academic funding bodies is mixed. I tried to access a large sample of resource enhancement projects funded by AHRC. There are some splendid and easily accessible sites, including Old Bailey Online, Fine Rolls of Henry III and the Nottingham University Place Name Tool and there are many other excellent ones. Sadly a number (5-10%) take you to broken links or cannot be tracked down. A few just do not understand democratic access. You can

only use the Newcastle Electronic Corpus of Tyneside English if you can demonstrate a bona fide interest. Hard to know what a non-bona fide interest might be and anyway you can hear a pretty good corpus of Tyneside English in the streets outside the university. You need to have been authorised by the University of Edinburgh to access the Calum Maclean collection of Scottish Gaelic folktales, but it is not clear how you gain such authority. For an amateur user, accessing the UK Data Archive poses real difficulties. You have to register – but the Archive makes it virtually impossible for non-academic users to sign on since you have to disclose your academic discipline. This poses a dilemma because non-academic users have *interests* not *disciplines.* If they decide to take the easy way out and pretend that they are doing a catering course, then the Archive requires them to fill in a box so that the Archive can ascertain whether the data they hold are suitable for the intended use.

Contrast that with the recent statement by the Mellon Foundation: *There are important public policy reasons for ensuring the broad reach of the humanities, and many of the library and scholarly resources and publications that the Mellon Foundation has supported are accessible and useful to a wide range of people from advanced scholars to students and teachers in kindergarten through 12th grade and the general public.* I have never seen a better definition of democratisation of access.

So, then a revolution in democratic access to records? I would say yes – but with a few pockets of resistance still to be mopped up.

So far, I have talked in very general terms about digitised records and have treated all digital collections as pretty much the same. In fact there is a very wide range of approaches from, at one end of the scale the "pile 'em high sell 'em cheap" approach of the census merchants who provide relatively unsophisticated approaches to assembling and making available collections of records. At the other end of the scale are some really sophisticated resources. The Oxford-based Electronic Enlightenment Project has a sales pitch of *Scholarship with added value* and says of itself EE *is not simply an "electronic bookshelf" of isolated texts but a* ***network of interconnected documents****, allowing you to see the complex web of personal relationships in the early modern period and the making of the modern world*. And at a cost of $2.7 million dollars to Mellon, so it might be.

One of the very real possibilities of this new world is that it is possible to use technology to create digital scholarly editions of texts. To quote Mellon again, *digital technology allows clear and elegant presentation of: variorum editions, which are relatively cumbersome to represent in print; multimedia editions of audio and visual as well as textual evidence; "editions as archives," which include facsimiles of original materials along with edited versions; and "editions of editions," which aggregate previously published editions of primary source materials to produce new and unique views of the evidence.* How long before the Royal Historical Society produces its first virtual Camden Volume? Oh well.

Some scholars in the field of classics have begun to describe fourth generation collections. These would include images of all source writings, whether these are on paper, stone or any other medium. They would be linked to XML transcriptions of printed versions of the writings and they may have advanced structural and basic semantic mark-up (e.g., careful tagging for each speaker in a play). They can use a small body of structured data — training sets, machine actionable dictionaries, linguistic databases, encyclopedias and gazetteers with heuristics for classification to find structure within the much larger body of content for which only OCR-generated text and catalogue level metadata is available. They also allow users to submit corrections and annotations and they should be able to determine how much weight to apply to various contributions, especially where these conflict. They posit a multi-layer system that can track contributions by both humans and automated systems, through different versions of the same texts. In addition, the systems will provide users with customised and personalised services and allow them to apply a range of technologies including text mining and visualization techniques.

I am not sure how these new technological approaches fit into our revolution. On the one hand, complex and sophisticated systems pose real problems of sustainability – I think there is an inverse link between the complexity of a system and its sustainability. It is unlikely that a library or archive would be able to take over a major high technology digital project without substantial external funding. It is interesting that Electronic Enlightenment is relying on subscriptions to develop its product.

The future is not going to be about producing digital scholarly editions as Mellon suggests: it is, as the classics people imply, going to be about machine-to-machine

communication. Scholars should be able to use computers to conduct analysis of a broad range of historical data from a range of sources. In order for this to happen, people developing new digital resources need to think about the use of linked data to describe them and of the provision of interfaces (APIs) to interrogate them. This is happening – but perhaps more slowly than it could. For an example of a technologically sophisticated approach which could be developed further, look at the Fine Rolls project. The other major technology which I believe to have possibilities for the future is the use of GIS and spatial data. There are limits to geo data since while it is great at describing space, it is not so good at dealing with time. But it does offer fantastic opportunities to offer an alternative view of history through the dynamic representation of time and place within culture. We are all looking forward to the Locating London's Past project which will create an intuitive GIS interface that will enable researchers to map and visualize textual and artefactual data relating to seventeenth and eighteenth-century London against a fully rasterised version of John Rocque's 1746 map of London and the first accurate modern OS map (1869-80) and Google maps. It will incorporate some of the future features I have described, including API methodology, to allow 'mash-ups' with modern datasets (geological, flooding, land use, etc). This in turn will create an environment in which additional external historical datasets and GIS enabled historical maps can be added.

And what about users. It seems to me that while the revolution may have brought many benefits, like most revolutions it has caused huge problems for ordinary citizens. I think that users face a number of difficulties which are magnified by the sheer wealth of the material which is available. One problem is the Google issue. If you are interested in a particular subject and try simply Googling it, there is a chance that you will be taken to a resource without any explanation as to its context and research value. Even going through a normal research channel you will soon find yourself in a maze of complex and difficult material. I have spent a lot of time reading eighteenth century newspapers online and I still find much of their content baffling – because what I am reading was written for a sophisticated contemporary audience who would have understood the all the mysterious allusions which puzzle me. When I started research a long time ago, I heard various urban legends about people who had wasted huge amounts of time and mental effort pursuing false trails – basing an argument on a known forgery or spending ages analysing financial records without understanding the basis on which they were

created. How many more opportunities are there for these sorts of errors in the digital world.

I am not sure whether researchers have the IT skills necessary to fully exploit this materials in new ways. The huge corpus of digital data means that new research questions can be asked and new answers sought, but this requires good technological skills. Researchers don't necessarily need to be able to write Python but they must have a good idea of the capabilities of data mining, linked data or geo-referencing material or whatever. So, as Matthew Davies said at the Gerald Aylmer Seminar – we need some form of training in digital literacy for the next generation of postgraduate students.

In terms of my original question, we are at the start of the digital revolution as far as the users of historical data are concerned. Partly it is a generational issue – how many professors of history know what Gate software does? Partly it is a skills problem and partly it is that the picture keeps on changing – I spent years studying relational databases and when I finally got it, object-oriented systems came along. Partly (and whisper this quietly) it is because IT training and education in the UK is not quite what it should be.

The final issue is that given the volume of digital resources which are now available online, most of which exist in stove-pipes of their own, there is an urgent need to demonstrate their scholarly value and to develop technologies which foster the aggregation of collections. I am convinced that somewhere in the 155 projects supported by the New Opportunities Fund, the 200-plus resources digitised by AHRC, the many records digitised by JISC or the commercial bodies, there is a document which is the key to my current research question. But I can't find it. I know it is not in Old Bailey Online or in a Vision of Britain, so where can it be? In 2008, the Mellon Foundation said that *given the depth and coherence of most of these [digital humanities] collections, funding priority in the Scholarly Communications program will shift from building the resources to activities that demonstrate and enhance their scholarly value and that foster aggregation of collections and the development of shared technology platforms in order to enhance sustainability*. It is interesting that the providers of Family History resources learned this lesson early on. 82% of family historians use subscription or pay-per-view sites – all of which provide a familiar and

safe way to conduct an aggregated search across a range of different types of records. As I said earlier, the New Opportunities Fund gave money to 155 separate projects, all of which were resolutely siloed. Now, through the wisdom of the IHR and the Connected Histories project, the most successful of those NOF funded projects, Old Bailey Online will be searchable along with a range of other significant resources and it is the vision of Connected Histories and the energy and determination of those who brought it to life which we are here to celebrate tonight.