IFLA WLIC 2013
SINGAPORE

Submitted on: July 19, 2013

born.digital@british.library: the opportunities and challenges of implementing a digital collection development strategy

Caroline Brazier
Director of Collections, British Library
London, United Kingdom

Abstract:

In this paper I will describe the recent evolution of the British Library's collection development strategy and policies, to show how one of the largest research library collections in the world is making the transition to digital collecting. I will discuss how we are changing what we collect in core areas such as commercial publishing, new web-based media and heritage collections. In doing so, I will consider how these changes are shaping our thinking about the future of services, and how they affect our approach to our key strategic partnerships. Finally, I will discuss how the move to digital leads to some fundamental changes in the knowledge and skills demanded of our staff.

Introduction

The British Library is the national library of the United Kingdom. Our core purpose is to make our intellectual heritage accessible to everyone, for research, inspiration and enjoyment. In support of this purpose, we build, curate and preserve the UK's national collection of published, written and digital content. We make the collections accessible to anyone who wants to do research through a wide range of library services, and also to the general public through a rich programme of exhibitions and events. Increasingly audiences expect to engage with our collections online, and we are evolving our collection development strategy to support this changing world.

In 2013 the British Library has published its most recent review of its Content Strategy. *From Stored Knowledge to Smart Knowledge: The British Library's Content Strategy 2013-2015* (www.bl.uk/aboutus/stratpolprog/contstrat) sets out our thinking on the changing landscape for building library collections in an increasingly 'born digital' world, and the way we are approaching it.

Collection development principles

We started out by revisiting the principles which have underpinned our collection development for much of the past decade, and which continue to inform much of our

selection and collection development work. While we had recognised the increasing importance of digital content in earlier reviews in 2005 and 2006, we had not fully addressed the challenges of moving from collecting physical materials to digital in all parts of our collection. In the current review of our principles, we have remained committed to the principle of using the structure of academic disciplines and subjects as the best way to communicate our priorities to our main user groups. So for digital collecting, as for our physical collections, we still focus on three main subject-based categories: Arts and Humanities; Science, Technology and Medicine; and Social Sciences.

In addition to the subject-based approach, another primary focus is on collecting materials in particular formats, and this remains a major channel for developing our collection, particularly where we are responsible for UK national collections such as sound or newspapers, or where we hold preeminent world-class collections such as maps or manuscripts. It is a principle to continue collecting in all these traditional formats as they evolve in the digital environment. However, we must also make new commitments to new types of digital 'formats' if we are serious about the principle of format-based collecting in the digital environment. Websites, social networking and news media channels are only some of the new types of digital format collecting we are doing. I will say more about this later in the paper.

The last major principle to highlight here is that in selecting content we will consider not only long-term collecting but also the crucial issue of access. For our digital collections to be relevant to 'born-digital' generations of researchers, we must take account of access and service models, and we must do this at the point of selection. Although the legal and contractual rules within which we operate will not allow online access to all content, we need to maximise such access where we can, and explain clearly why we cannot in other cases.

A changing landscape: legal and policy issues affecting collection development at the British Library

2013 will be seen as a landmark year for the British Library and the other five legal deposit libraries of the UK. It was on 6 April this year that the new UK legal deposit regulations for non-print publications finally came into effect. The journey to this point had started almost 20 years before with the writing of the first internal document in the British Library to address a growing number of questions. Should the Library collect new digital formats? What should we collect? How should we collect it? This document warned of problems for the future comprehensiveness of the national published archive if the Library was unable to start collecting the growing number of publications appearing on disc and CD. After a ten-year period of research, discussion with the publishing industry and negotiation with government, the Legal Deposit Libraries Act was passed by UK government in 2003. This Act established the principle of digital legal deposit collecting but required further work on more detailed regulations, involving both libraries and publishers. It was not until 2013 that these regulations were finally agreed and implemented, allowing the six UK legal deposit libraries to start collecting UK digital publications at scale and, most importantly, to start archiving the UK web domain.

In parallel with the campaign to achieve digital legal deposit, we have also seen more than a decade of campaigning internationally for increased open public access to research findings based on public funding. In the UK, the Finch report (Finch, 2012. Accessibility, sustainability, excellence: how to expand access to research publications. Report of the Working Group on Expanding Access to Published Research Findings) proposed new open-

access publishing models for research publications to achieve better, faster access for anyone who wants to read or use them and help reap the full social, economic and cultural benefits that can come from research. This report was the catalyst to a significant change in the culture of research funding and led to various mandates for open access from organisations which fund academic research in the UK. Between them these have radically changed the publishing models for research articles in particular. For the British Library, these cultural and operational shifts in publishing models will inevitably change the way we provide access to research articles through our services and affect our thinking about the 'value' of journal subscriptions. Open-access models for research e-monographs are also being considered, and these are likely to lead to equally far-reaching changes in collection development.

Copyright frameworks within which we build and use digital collections are also changing. In the analogue world, library collections were consumed in a relatively passive environment. Technology limited what you could do. Today, however, we live in a world not only of digital 'access' but of active digital 're-use'. Copyright law is very relevant to many of the things libraries want to do with post-1870 collections and, as a profession and as individual institutions, we constantly try to influence the development of the legislation in a positive direction. The 1990s and 2000s can be viewed as a period of copyright expansionism, during which global law essentially made sure that the internet was safe for publishers. We may now be seeing a counter-movement, with the World Intellectual Property Organisation (WIPO) considering an international treaty for libraries and researchers, and new laws to facilitate mass digitisation and to deal with orphan works in Europe and the UK (http://www.ifla.org/publications/are-digital-laws-making-or-breaking-digital-libraries-0). Legal exceptions in support of data and text mining are also being considered as governments start to see potential for growth in data-driven innovation and big data.

In a world of physical collections, what can be done with the content is limited, but the process of providing access is straightforward. In a digital world, there may be a gap between what is possible and what is permitted. The legal, contractual and regulatory frameworks thus have a major influence on what libraries can do with their collections in a digital world. Their influence over the choices we make in building our digital collections is already enormous, and in the future they will become even more central in our decision-making. Some libraries may choose not to collect certain types of materials if they are not able to give the kind of access services and re-use functionality their customers demand. The cost of investment in digital content may not be justified, unless today's users get the benefits of services which meet their needs.

For national libraries this choice is often not so clear-cut. While the services we can offer our users today are very important (and they often influence the perceptions held by politicians and policy makers about us) we are also collecting on behalf of future generations of our citizens. We must make collection development decisions not only to meet the needs of today but to anticipate those of tomorrow. This means that we must be able to communicate clearly what it means to be a national library in a digital age.

Building a national collection in a digital age

What does it mean to build a national collection in a digital age? In considering such a question, we are likely to focus on the concept of the 'digital age', but there are two other important words in the question, and we need to take a step back – to consider what we mean by 'national', and what we mean by 'collection', in this context.

Defining 'national'

The web is global, and digital content does not respect national boundaries. Much of it flows freely around the world. And where content sits behind barriers such as passwords or pay-walls it is generally who you are or what you are, not where you are, that prevents access. Your failure to reach the content may be because you are not a member of the right group, or because you are too poor to pay for it, but it is unlikely to be because you are in Andorra or Zimbabwe.

And more and more content producers are working on an international or global scale, which can make it difficult to answer the question 'In what country is this book published?' Under which legal deposit regime do you collect an e-book about an American artist, written by a British author, commissioned by a UK publisher but out of their Australian offices, with the editing and production subcontracted to India who in turn subcontract to editors around the planet?

Under the recently developed UK legal regulations, an online publication is treated as 'published in the UK' if it is made available from a website with a domain name which relates to the UK or a place within the UK, or if it has been made available by a person whose activities relating to the creation or publishing of the work took place within the UK. This means that for national collecting, identification of place from domain names becomes increasingly important. In many other cases, only discussion between library and publisher can establish where the substantial part of creating the work took place.

Defining 'collection'

What makes a digital collection? In the world of analogue collections, we understood what it meant for something to be 'in' a library collection. When items were selected to be added to our collections, we brought them into the library, added data about them to our catalogues, stored them on our shelves, and preserved and conserved them for as long as they were core to our sense of purpose (which usually means 'forever' in the case of national libraries). But the nature of digital information and materials means we can now 'collect' in very different ways. As well as ingesting digital copies of publications into our own servers, we can also connect to information resources and publications for which we have obtained access rights through licences, membership or partnership arrangements. Connecting to content becomes as important as collecting it. And increasingly, recommending particular open-access content by choosing to link to it will become as important a part of collection development as selection was in the analogue world. As well as stewardship of digital content under our direct control, libraries will develop their 'collections' through expanding access rights on behalf of users, and also expand their users' trust in the quality and appropriateness of what we choose to link to and recommend. Access rights are replacing physical ownership as the fundamental definition of being 'in' a library collection.

National libraries are collecting to ensure future, as well as contemporary, access to national memory. So how confident are we that we can rely on others to preserve the connections to content we consider of national importance for the long term? What challenge does digital collection development have for the long-term preservation of commercial research materials? An increasing number of publishers and libraries are actively participating in digital preservation systems through organisations such as Portico (www.portico.org) or CLOCKSS (www.clockss.org). However, there is as yet no systematic guarantee of comprehensive collecting and preservation working together in an integrated way to guarantee access for future generations. The services are robust, but nonetheless there are some concerns over long-term viability.

The British Library, like many national libraries, feels that it must still ingest digital content of long-term importance to the national collection, to guarantee long-term access for future generations. This is perhaps the most challenging area of digital collection management, but it also offers new opportunities for shared services and future collaboration. In particular, it opens the door to greater international collaboration. In the digital age, each national library not only has the opportunity to continue to underpin its national collection development within its own country, but it can also play its part in guaranteeing long-term international access through shared digital collection development and management. Through international collaborations, we need to rethink the principles of UBC (Universal Bibliographic Control) and UAP (Universal Availability of Publications) for the digital age.

Digital collection development in practice at the British Library

I would like to explore what this means in three different areas of digital collection development – collecting commercial publishing, collecting new born-digital 'carriers' of information, and collecting digital heritage materials.

Collecting commercial digital publications

The British Library has traditionally spent a considerable proportion (c.20%) of its budget to acquire research publications, mostly internationally published academic journals and monographs, to enrich our collections beyond UK legal deposit and support our range of information services for UK and international researchers. As well as access through our reading rooms in London, we have also offered the largest document supply service in the world.

Over the past decade we have bought international research publications in digital format as much as possible, to improve speed of access and to help us deliver cost savings in collection management and storage. However, future expansion of this policy is threatened by the rising costs and also the position of publishers and copyright licensing agencies in what they are prepared to license. Researchers increasingly expect access wherever they are, but we have found it difficult – and I suspect it is the same for other large national libraries – to license online access outside of our reading rooms for our registered users. It has been impossible to consider negotiating national licences, as publishers are unwilling to consider any national licence that would bring them less revenue than the sum of the individual institutional licences in the country. With a large population (60 million) and large higher education and industrial library sectors in the UK, the cost of a national licence would be prohibitive. As with most national libraries, membership of the British Library is open to everyone, so publishers see a general open licence as potentially undermining revenue from other sources.

But can the difficulty of negotiating national-level licences also provide an opportunity to reconsider what should be in the national collection? As the higher education sector develops new student financing models, as the British Library's budgets are reduced by cuts in government grants, and as licensing and copyright restrictions prevent us meeting user expectations, we must question whether it is the best use of limited funds for national collection development to duplicate non-UK content which is already widely purchased by the higher education sector. Many of our registered users are also members of academic institutions, as academic staff or as postgraduate or undergraduate students, and with remote authentication they can access it anywhere, including in our reading rooms.

Our document supply services have also been facing declining demand because of the widespread availability of many journal titles through the 'big deals'. This is coupled with increasing competition from new commercial document supply services, which do not collect and manage content but simply offer aggregated discovery services and act as resellers for publishers. Moreover, many publishers now offer their own online ordering services to their current and backfile content. Against the background of these developments, it is increasingly difficult to justify collecting as much non-UK published e-content for these services.

As this non-UK e-content is not part of our legal deposit remit, and as much of it is already covered by international and community-based shared digital preservation services, we no longer have to feel responsible for collecting international titles to preserve them. Digital content therefore enables us to consider taking a more radical approach and questioning whether we can provide services in different ways. If we can guarantee UK researchers access when they need it, without having to collect as comprehensively, we will be prepared to take a radically different approach.

Another new question in digital collection development is what the role of the national library should be in collecting and preserving the increasing amount of research article literature being published by various open-access routes. The rise in open-access publishing of research findings that are the result of publicly funded research is the culmination of a decade and more of campaigning. In theory, there seems an obvious role for national libraries to collect and guarantee preservation and long-term access to these important outputs of national investment in research. In practice, the multiple models of open-access publishing do not make this easy to achieve. Without consistent metadata standards and agreement between the 'owners' of the open-access channels, including commercial e-journal publishers, open-access journals, institutional and subject repositories, it is technically and operationally challenging to offer aggregated collecting, discovery and preservation services. Agreeing an appropriate role for national libraries in long-term collecting of open-access content is one of the major challenges we will face over the next decade.

We have also undertaken a major policy review of our approach to collecting newspapers and related media. The British Library has built one of the largest newspaper collections in the world, including both UK and international newspapers. We have also collected radio news broadcasts and, more recently, digital news broadcasts available in the UK, as part of our Sound Archive. Under the new legal deposit regulations, we have reviewed our collection development strategy for all news formats. We have decided to break down these traditional barriers and adopt a more integrated approach to news media in all formats, including not only newspapers but also radio, broadcast and web-based news. This new approach modernises our collection development strategy and ensures it reflects the changing information-seeking and access behaviours we see around us.

As well as rethinking our collecting policies for contemporary commercial publishing, we are also facing a challenge with access to retrospective commercial publishing. Mass digitisation of backlists of journals, newspapers and similar content types by publishers, media companies and third-party companies is introducing new ways for the public to access retrospective collections. Digital technologies have offered content owners new opportunities to offer access services and exploit their content and archives in new ways. A major challenge to our role as the national library is that we are no longer the only UK organisation who can capture, preserve and make available our country's published heritage. We need to explore new partnerships to turn it into a new opportunity too.

Collecting new born-digital formats

Digital collection development for commercial publications may offer us opportunities to rethink and adapt our collecting policies. But the real 'prize' of digital is the opportunity to build totally new types of collection from scratch.

Web harvesting is an essential area of collecting, if national collections are to reflect the way we live and communicate today, for the benefit of future generations. It also presents many challenges, both technically and operationally, because we must collect at such scale. As part of the work to support regulations for UK legal deposit, in 2013 we estimate there are 4.8 million websites in the UK domain. The British Library, on behalf of the UK legal deposit libraries, plans to crawl the entire .uk domain and other sites of UK relevance once or twice a year. This will give breadth to a future web archive, but not necessarily the depth required to cover rapidly changing sites. In addition to machine crawls, we plan to do selective or 'curated' crawls on a more frequent basis, where individual sites or a collection of related sites which are judged to be important to national events, for cultural or subject reasons, and which change their content frequently, are selected for more frequent capture. To be useful to researchers in the future, collection development for websites must also be able to respond quickly to rapidly emerging events, including natural disasters and political events such as the Arab Spring, where selection of sites cannot be planned in advance. This type of 'rapid response' collecting requires a new approach to selection from curators and subject specialists, and a significant shift from traditional collection development activities.

Current limitations on access to the legal deposit collection also affect collection development policy. As legal deposit content can only be viewed on the premises of the six legal deposit libraries, and as there are sites where we feel offsite access would be valuable to researchers, we are planning to run a parallel selection process for sites where we will need to request the site owners' permissions not only to harvest content but also to make it available. This three-tier process – broad domain crawl, high-frequency selective crawl and permissions-based archiving – will be important in ensuring that we can build a web archive that is increasingly comprehensive but responsive to researchers' short- to medium-term needs. A more challenging question is how sustainable such a complex mix of automated crawl and curatorial selection will be in the future, as the scale of the UK domain grows.

Collecting born-digital heritage

Special collections and rare or unique heritage materials have been amongst the most iconic areas of collecting for many research libraries in the analogue world. The key challenge over the past decade has been the digitisation of many of these collections to make them accessible in a digital age. But increasingly, national libraries which collect archival and heritage materials must address born-digital materials, which present a new and very different range of challenges.

Special-format collections require a particular focus, as the whole nature of a collection changes along with the expectations of researchers about types of access and what they should be able to do with digital equivalents. In collection areas such as maps, the replacement of physical maps with digital mapping data presents new challenges in large-scale data management. However, it also opens up new avenues for research and engagement, enabling researchers to add to the value and interpretation of the collections. The British Library's geo-referencing projects (www.bl.uk/maps) added new knowledge to the collection by crowdsourcing the overlaying of historic digitised maps with modern born-digital mapping.

Collecting digital equivalents of private papers and archives is an area which offers challenges on many levels. Personal papers, records and archives of leading individuals in any field of study has always been a mainstay of special collections because of the unique insight they give not only into the development of an individual's work but also into the wider context within which that individual's work developed. Like everyone else, such people now work in increasingly complex ways, sharing information about themselves and their work with wider audiences and through multiple channels. National libraries which collect personal papers are facing new types of challenge as our personal communications channels multiply and increase in scale. What should we collect from whom, when and how?

The question of what to collect in a digital environment may be the least of these challenges. Writers, artists, politicians and scientists may work fully in digital environments, but the nature of that work remains largely the same. There are obvious digital equivalents for many of the types of materials traditionally collected, so we collect the poet's emails instead of his letters, the scientist's digital laboratory notebooks instead of her physical notebooks, and the politician's private digital memoirs instead of written diaries. But as well as simple digital equivalents, libraries attempting to collect a complete digital 'life' also have to collect from many new channels of digital communications through which people express themselves. So we need to be able to collect individuals' websites, their blogs, their tweets and multiple social networking 'personas' as well as the truly private digital 'papers'.

There are many new challenges in building this type of digital collection. The first challenges are technical and financial, for we need to be able to respond quickly to collect new types of content, from new communications channels or new devices. Because of the relative speed with which new channels of expression and communication come along, and the relative fragility and impermanence of digital content within them, we need to start collecting much earlier from individuals than in the analogue world in order to ensure that important digital collection materials do not vanish or get altered. It is also important to understand the technical context within which content was originally created. It is far easier to recreate the context of different word-processing or email systems while they are still current, rather than to rely on the digital archaeology required to do it twenty years after the software was last publicly available.

The need for intervention with individual subjects and their 'papers' as early as possible is also changing the nature of selecting from whom to collect. In the analogue world, relationships with potential subjects and/or their families could be developed over many years. Collecting libraries traditionally waited until they were sure that an individual's work merited the collecting of his or her papers, as the nature of the collection materials themselves would remain unchanged. In the digital world we do not have that luxury. The development of relationships with potential subjects, and the initiation of collecting, must begin much earlier in a subject's career cycle to avoid much valuable material falling into a digital 'black hole'.

A third question is whose material is it anyway? The ownership of many strands of digital content is not always clear-cut. Subjects may assume they own what they write in blogs or tweets or social networking channels, but this may not always be the case. They may have inadvertently given away rights by not reading the small print of licence and contract terms. Libraries wishing to collect personal digital documents have a new set of copyright, privacy and intellectual property rights issues to contend with.

A major question for the international national library community is whether we can share responsibility for any of this type of collecting. The technical preservation challenges posed by individual archives are only beginning to be understood, and it is likely that the cost

implications of this type of development may in future limit the number of national libraries that will be able to collect such material. Shared solutions and service models are one way forward. Can we find ways to integrate national collections and their curation with the centralised collections of aggregators such as the Library of Congress in the case of Twitter (Library of Congress 2013)? Can special collecting become shared collecting, with some institutions acting as gatekeepers of privacy and rights?

Developing staff skills to support digital collecting

A major challenge in implementing digital collection development strategies lies in ensuring that the staff in the Library are equipped with the necessary skills and capabilities to select, collect, manage and curate digital content.

In the British Library, we have been working on this issue for several years now. As a library which works to support researchers in all disciplines and subjects, we have seen major shifts in researcher behaviours, in publishing, and in the scholarly communication chain. The impact of new technologies and 'big data' on libraries and librarians is most advanced in e-science and e-social science. However, the rapid evolution of digital humanities research, based in part on the new potential from mass digitisation of textual materials, is fast catching up in its impact on the way libraries must work.

At the British Library, we have established a small team of 'digital curators' who have the responsibility to explore this new world and recommend ways to make the necessary skills mainstream in our curators and subject librarians. Based on their work, we have developed a Digital Scholarship Training Programme, which is a two-year programme consisting of 15 one-day courses which look at a range of the key concepts, methods and tools that define today's digital scholarship practice. The content of the course was designed in consultation with leading experts in digital humanities, as well as major partners in the higher education, cultural heritage and IT sectors.

We start by offering our staff a series of short updates in some of the 'basics' of working with researchers and collections in a digital environment. These cover topics such as use of social networking, updating presentation skills, basics of working with digital objects and images, basics of the web and programming, and an update on metadata for electronic resources including Dublin Core, METS, MODS and XML.

This foundation is intended to make sure that everyone can get the most out of the main courses, which explore questions such as

- How are digital scholarship and digital collections changing the nature of research?
- What are the lifecycle issues involved in managing digitised and digital collections, from selection through to long-term preservation?
- How can we design new access and re-use services for digital content, taking into account copyright, licensing and intellectual property rights?
- What is best practice in crowdsourcing in cultural heritage and memory organisations, to maintain authority and trust?
- What are the new ways to present digital content through text encoding, data visualisation and geo-referencing?
- What new opportunities exist for information integration and sharing through mash-ups, APIs and the semantic web?

We are just starting the second year of the course, and so far the feedback has been very promising. It has helped to demystify many of the new technologies which we must be able to use in order to work with new generations of researchers. It has raised confidence in many staff about their ability to work effectively with many new types of digital content and digital services. As new ideas flow from staff, because of their increased awareness and confidence, it has helped us become more creative as an organisation and better able to prioritise digital developments against a background of on-going budget reductions.

Conclusions

In this short review I have examined some of the principles of collection development in a digital environment, and some aspects of the changing environment which affects the choices we have. In terms of practical responses, I have focussed on what is happening in the British Library. But I hope that much of this has relevance to other national libraries across the world.

In conclusion, I should like to single out two specific points. First, we must be open to radically new approaches to what we collect and what we can enable our users to do with it. There are enormous opportunities to offer new types of collections and new ways of using content. Secondly, I hope that we can grasp the challenges presented by a rapidly changing and increasingly digital environment not as a threat to our roles as national libraries but as a chance to rethink the potential of what national libraries can do. And we must see this potential not just within the boundaries of our own countries, but also have a renewed vision to what we can achieve together as an international networked community.

Caroline Brazier

Director of Collections

British Library