Identifying Digital Preservation Requirements: Digital Preservation Strategy and Collection Profiling at the British Library

Michael Day
The British Library
96 Euston Road, London NW1 2DB
United Kingdom
+44 (0)843 2081144 x 3364
Michael.Day@bl.uk

Ann MacDonald
University of Kent
Canterbury, Kent, CT2 7NZ
United Kingdom

Akiko Kimura
The British Library
96 Euston Road, London NW1 2DB
United Kingdom
+44 (0)20 7412 7214
Akiko.Kimura@bl.uk

Maureen Pennock
The British Library
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom
+44 (0)1937 546302
Maureen.Pennock@bl.uk

ABSTRACT

The British Library is increasingly a digital library. Over past decades, it has built up significant collections of digital content covering a very wide range of content types. In addition to the increasing amounts of digital content acquired by purchase or donation, the Library and its partners have also invested heavily in the digitization of selected collection content, helping to create large collections of certain types of content (e.g., newspapers, out-of-copyright books, and sound). Most recently, the extension of legal deposit provisions to non-print works in 2013 has meant that the British Library - working in conjunction with the other UK legal deposit libraries - has begun to collect new categories of digital content, including periodic harvests of the UK Web domain. In order to support this, the Library has also invested heavily in developing scalable infrastructures for the acquisition, storage and management of large amounts of digital content. The British Library Digital Preservation Strategy, 2013-2016 is focused on the embedding of digital sustainability as an organizational principle across the Library and to help manage preservation risks and challenges across all digital collection content lifecycles. This practice paper describes work being undertaken by the Digital Preservation Team at the British Library to develop content profiles of high-level digital collections that will support the implementation of the strategy, in particular for the capture of long-term preservation requirements.

General Terms

strategic environment, preservation strategies and workflows, case studies and best practice

Keywords

digital preservation, collection content profiling, preservation planning, institutional contexts of preservation

1. INTRODUCTION

This paper describes work being undertaken by the Digital Preservation Team at the British Library to develop a content profiling framework for high-level digital collections that will help support the capture of long-term preservation requirements. The resulting collection profiles are short human-readable documents that document and contextualize collections that then can be used as part of the preservation planning process.

This paper will follow the following structure. After a section describing the digital preservation context of the British Library, section 3 will outline related work in the areas of preservation planning, content characterization and profiling, the capture of preservation intent, and some approaches to institution-level assessment. Section 4 will then describe in more detail: 1) challenges around the identification of high-level digital collections at the British Library, and 2) the development of the initial collection profile framework. Section 5 provides some conclusions and pointers to future work.

2. THE BRITISH LIBRARY CONTEXT

The British Library is the UK’s national library; its role is defined in legislation as “a national centre for reference, study and bibliographical and other information services, in relation both to scientific and technological matters and to the humanities” [British Library Act 1972].

2.1 Legal Deposit

As a legal deposit library, the British Library has the right to receive a copy of printed content published in the UK (including books, newspapers, printed music and maps) as well as - since April 2013 - certain kinds of non-print content. For printed materials, this obligation has existed in English law since the seventeenth century. Primary legislation supporting the extension of legal deposit to non-print items in the UK was passed in 2003.

After a decade of planning and negotiation, official regulations came into force on the 6th April 2013 [Legal Deposit Libraries (Non-Print Works) Regulations 2013]. This, for the first time, enabled the British Library and the other UK copyright libraries to claim certain classes of non-print content under legal deposit provisions and make it available to on-site users [Gibby and Brazier 2012].

This has included the scaling-up of the Library's existing Web archiving activities to include a periodic capture of the entire UK Web Domain, the first of which (running from April to June 2013) captured 31TB of compressed data [Webster 2013]. It has also led to the development of specialised ingest workflows for the capture of other kinds of published content, including e-journals and e-books.

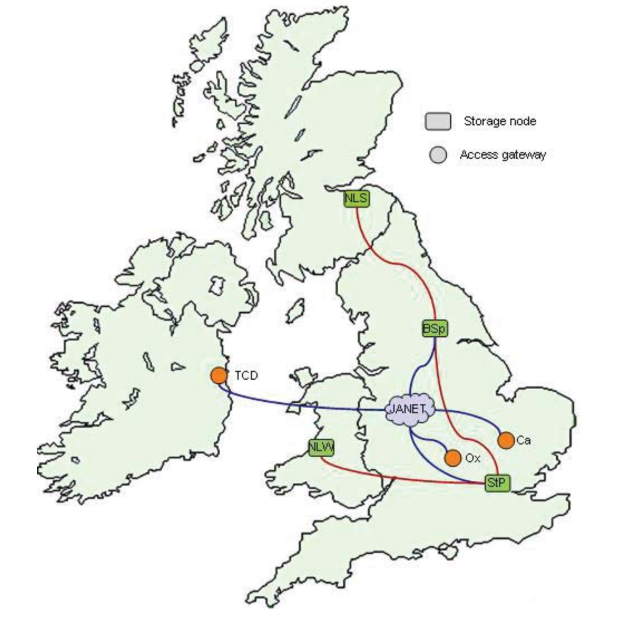
2.2 Infrastructures

In order to scale-up its technical infrastructure, the British Library and the other UK Legal Deposit Libraries have invested heavily in developing scalable solutions to the acquisition, storage and management of very large amounts of digital content. The resulting Digital Library System (DLS) has been described as a "single location to ingest, store, preserve, manage, discover and provide controlled access to digital content assets" [Fleming 2011]. While designed as an integrated storage system, it has been implemented in a highly distributed way, with content replicated in four storage nodes (based in London, Boston Spa, Edinburgh and Aberystwyth) with additional access gateways at the university-based legal deposit libraries (Figure 1).

Some features of DLS have been described in an APARSEN project deliverable [APARSEN 2013]. Ingest takes place at either of the British Library's sites, with different ingest streams defined for different types of content, e.g. e-journals, digitized newspapers, or web archive content. All objects have a signature file, which includes a hash value and timestamp, and content is automatically replicated on all four storage nodes after ingest. The system assumes that in a large-scale storage system, some bit-loss is inevitable. DLS has, therefore, been designed to be self-checking and self-healing; there are periodic integrity checks, and "if an object is found to be damaged, it is replaced by a good copy from another node" [APARSEN 2013]. DLS is designed to be scalable and vendor-independent, using commodity hardware that can be added to as required.

2.3 Strategy

At the same time, the Library has begun to try to understand what might be meant by a "national collection" in a digital age. It has been widely recognized that the meaning of traditional concepts of "collection" (and therefore "collections management") have changed significantly in the digital era, e.g. being focused much less on 'tangible' content held and managed locally and more on providing access to content held elsewhere [Corrall 2011; Corrall and Roberts 2012]. In this environment, a great deal of attention needs to be given to access rights. For example, Brazier has commented that "access rights are replacing physical ownership as the fundamental definition of being 'in' a library collection" [Brazier 2013]. This shift is also seen in the British Library's Content Strategy, 2013-2015. While recognizing the continuing significance of collecting activity, e.g. through legal deposit, voluntary deposit and donation, the strategy states that outside of this, "the Library will prefer to connect to content, except in circumstances where the connection is not technically feasible or when we wish to hold and preserve the materials for the long term" [British Library 2013a]. Despite this, the logic of Non Print Legal Deposit, Web domain harvesting, and the Library's ongoing digitisation partnerships mean that the amount of digital content that will require long-term preservation is growing at an extremely rapid rate.



Storage nodes:

- British Library, St Pancras (STP)
- British Library, Boston Spa (BSP)
- National Library of Wales (NLW)
- National Library of Scotland (NLS)

Access gateways:

- Bodleian Library, Oxford (Ox)
- Cambridge University Library (Ca)
- Trinity College Library, Dublin (TCD)

Figure 1. DLS Storage Nodes (Source: APARSEN 2013)

When all of this is taken into account, it is clear that the British Library is increasingly becoming a digital library. The British Library's Digital Preservation Strategy, 2013-2016 starts from the assumption that it is the Library's responsibility to preserve and make available this content to current and future users, while noting, however, that "preservation of digital content is not straightforward" in that it "requires action and intervention throughout the lifecycle, far earlier and more frequently than" with physical collections. The strategy, which was approved in March 2013, outlines four main strategic priorities [British Library 2013b], i.e. to:

- *Ensure [the Library's] digital repository can store and preserve [...] collections for the long term;*
- *Manage the risks and challenges associated with digital preservation throughout the digital collection content lifecycle;*

- *Embed digital sustainability as an organisational principle for digital library planning and development;*
- *Benefit from collaboration with other national and international institutions on digital preservation initiatives.*

At least three of these priorities depend upon there being adequate knowledge of the British Library's digital collections, e.g. for being able to establish and invoke suitable preservation plans, for monitoring the wider technical environment (preservation watch), or for building awareness of digital preservation issues amongst Library colleagues and (ultimately) its users. A useful first step appeared to be to work with curators and other content specialists to develop descriptive profiles of the Library's high-level digital collection areas, with the aim of capturing key knowledge about the collections and their specific preservation requirements.

The British Library's Digital Preservation Team has, for the very first time therefore, begun to develop content profiles for the Library's high-level digital collection types. It is intended that these will help provide the opportunity to build conversations with curators and content specialists on identifying specific preservation requirements. This has a number of benefits:

- The massive scale of content held by the British Library means that collection profiling is a crucial part of preservation planning, supporting the identification of preservation requirements, and the tools necessary to facilitate these.
- Collection profiling opens a forum on which collection stakeholders, the people who make decisions at different lifecycle stages, can express challenges faced by specific content types. This should help the development of a shared understanding of digital preservation requirements from both curatorial and technical perspectives.
- Corporate understanding of the collections held by the British Library is enriched through the sharing of collection information, between the departments which make collection decisions. This acts as a platform on which to build sustainable preservation development.

3. RELATED WORK

The British Library's collection profiles are intended to support the planning of digital preservation activities across different content lifecycles. It, therefore, builds on previous work focused on the assessment of content, including the use of decision support tools for preservation planning, the use of tools and registries for content profiling or characterization, as well as more direct attempts to capture curatorial 'intent' for specific collections. There is also a link to institution-level assessment (e.g. repository audit) in that audit tools and maturity models could potentially also be applied at collection or ingest work stream level. The work is also related to ongoing research on defining the significant properties (or characteristics) of digital objects, not least in taking account of how significance may be understood differently by the various stakeholders involved in the preservation process, including creators, custodians and consumers [Knight and Pennock 2009; Dappert and Farquhar 2009].

This section will outline some related work on the assessment of collections for digital preservation, focusing on preservation planning decision-support tools (e.g. Plato), technical content characterization tools, the capture of preservation intent, and assessments at the institution or repository level.

3.1 Preservation planning

The OAIS Model defines a Preservation Planning Functional Entity that "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete" [ISO 14721:2012; CCSDS 650.0-M-2 2012]. It also provides some specific examples of what functions might be required:

Preservation Planning functions include evaluating the contents of the Archive and periodically recommending archival information updates, recommending the migration of current Archive holdings, developing recommendations for Archive standards and policies, providing periodic risk analysis reports, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base. […] Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

It is clear from this that preservation planning is a critical component of any digital preservation strategy.

One attempt to develop a structured approach to preservation planning is the Plato decision-support tool developed as part of the Planets (Preservation and Long-term Access through Networked Services) project [Becker et al 2008; Becker et al 2009]. Plato provides a methodology and a software tool to support the systematic capture of preservation requirements from various stakeholders and then to match these to potential preservation strategies for further analysis. The result is a recommendation that can form the basis of a preservation plan, which contains information on contexts as well as the evidence base underpinning the decision.

There have been various attempts made to integrate Plato with other digital preservation systems. For example, researchers from the KeepIt and Planets projects integrated Plato and other digital preservation tools with the ePrints repository software, creating plugins to ePrints that would support the development of preservation workflows, including the generation of preservation plans and action plans [Hitchcock et al 2010].

The SCAPE (Scalable Preservation Environments) project[1] has also been exploring how to integrate Plato with other digital preservation tools and services [May and Wilson 2014]. The project is specifically interested in enabling Plato to:

- Import information from external sources, e.g. from content profiles or from institutional policies.
- Integrate with other services, e.g. the SCAPE's Component Catalogue of tools or the Scout automated preservation watch service [Faria 2013]
- Incorporate planning functionality within repository systems, so that plans can be fed back for monitoring

In terms of SCAPE, the resulting Preservation Plans document collections, their institutional context, and the decision-making process that led to the selection of a particular preservation action. It also contains a Preservation Action Plan that contains all of the

[1] SCAPE project. Retrieved August 30, 2014 from http://www.scape-project.eu/

information necessary to apply the preservation action as well as an Executable Action Plan that can be deployed through a workflow management system (e.g. Taverna).

3.2 Content characterization

Preservation planning support tools like Plato depend upon there being accurate information about the file representation types (e.g. formats) present in a collection or repository. The scope of this has been outlined by Faria et al [2013]

Digital preservation starts by understanding what content a repository holds and what are the specific characteristics of that content. This process is supported by the characterization of content and allows a content owner to be aware of content volumes, characteristics, format distributions, and specific peculiarities such as digital rights management issues, complex content elements, or other preservation risks.

Several different tools and services have been developed to help with content identification, validation, and characterization, of which JHOVE (JSTOR/Harvard Object Validation Environment) and DROID (Digital Record Object Identification) are perhaps the most well-known. Characterization software like JHOVE, JHOVE2 or DROID can in turn be embedded into other tools. For example, the File Information Tool Set (FITS), originally created by Harvard University Library, combines a number of different open source tools – currently including JHOVE, DROID, Apache Tika, and the National Library of New Zealand Metadata Extractor – in order to support consistency of use across all tools and to produce standardized output metadata [McEwen and Goethals 2009].

It is obvious that with ever growing collections, characterization tools need to work at scale. One recent initiative has been c3po (Clever, Crafty, Content Profiling of Objects), which has produced a prototype software tool that produces content profiles of collections based on data generated by FITS that can be used for further analysis or visualization [Petrov and Becker 2012]. Tools like DROID and Apache Tika have also been used to analyze very large collections, e.g. Web archives, where there is considerable interest in the use of scalable characterization tools [Jackson 2012; Palmer 2014].

The British Library has an active interest in content characterization tools, not least through its involvement in the SCAPE project, one of whose objectives is enabling the large-scale characterization of digital objects [Van der Knijff 2011]. The British Library Digital Preservation Team's current work-plan also contains work-streams for file format assessment; tool assessment and preservation watch, all of which will involve some level of content characterization at a technical level.

3.3 Content profiling

While the technical aspects of content characterization remain important, the British Library's collection profiling activity described in this paper has primarily drawn its inspiration from other content profiling activities, i.e. those based on a structured dialogue with curators and other content specialists. As part of the collection profile development, a number of content-based profile initiatives were reviewed, in particular the Digital Content Reviews (DCR) for Life Cycle Management developed by MIT Libraries and the Data Curation Profiles developed by Purdue University Libraries.

Purdue's Data Curation Profiles are a tool for capturing basic information about research datasets in order to support their curation and reuse. The profile provides a framework (an interview structure) that can be used to gather information about datasets and their potential re-use. Once completed, profiles can help guide decision-making about the management of datasets as well as inform those providing research data management services of any specific requirements [Witt et al 2009]. The Data Curation Profile toolkit (an interviewers' manual/worksheet and user guide) has been made freely available, and the profiles have begun to be used in other initiatives, e.g. by Cornell University Library to help design the Datastar research data registry [Wright et al 2013]. While the Data Curation Profiles were probably too focused on one particular type of content to be useful for our immediate purposes, the general approach clearly demonstrated the benefits of using content profiles to support lifecycle management.

MIT Libraries' Digital Content Reviews for Life Cycle Management took a similar lifecycle-management view, but – more like the emerging British Library profiles - were primarily intended to help capture information about the implications of collecting certain types of digital content. The section headings are a mixture of generic (content overview, collection management, rights management) and those that follow the content lifecycle (acquisition, ingest, preservation planning, archival storage, long-term access) [MIT Libraries 2013].

3.4 Preservation intent

While these existing content profiles provided us with a basis for developing a draft framework for the British Library profile, another key inspiration was the National Library of Australia's work on identifying 'preservation intent' [Webb et al 2013]. As part of their approach to preservation planning, digital preservation specialists at the National Library of Australia have been concerned to talk to content specialists (collection managers, curators) in order to develop some 'plain-language' statements about "which collection materials, and which copies of collection materials, need to remain accessible for an extended period, and which ones can be discarded when no longer in use or when access to them becomes troublesome." Content specialists were also "asked to make broad statements clarifying what 'accessible' means by stating the priority elements that need to be re-presented in any future access for each kind of digital object type in their collections." This both becomes a means of ensuring that curators and other collection specialists take responsibility for deciding what will happen to collections and is essential for preservation planning. Webb et al [2013] write that "without it, we are left floundering between assumptions that every characteristic of every digital item has to be maintained forever (almost certainly an impossible expectation) and assumptions that it is good enough to store data safely and let future users worry about how to access it (almost certainly an inadequate response)." Capturing elements of preservation intent seemed vital for the success of the British Library's collection profiling activity.

3.5 Institution-level assessment

Other approaches to digital preservation assessment have been focused on higher levels of aggregation than collections. This includes well-established work on repository audit, where the main focus of attention has been on two interrelated standards:

- The Trustworthy Repositories Audit & Certification (TRAC) criteria and checklist published by the US Center for Research Libraries [2007]
- ISO 16363:2012 Audit and Certification of Trustworthy Digital Repositories [ISO 16363:2012].

Both provide a framework for the assessment of repositories based on three main categories: organizational infrastructure (including governance, structure and financial sustainability), digital object management, and infrastructure and security risk management.

These standards mainly focus on organization and infrastructure rather than collections, but some other approaches to institutional evaluation do have the potential to be able to inform the assessment at collection-level. This is particularly true of approaches based on maturity modelling, which include the Digital Preservation Capability Maturity Model, whose levels mainly focus on perceived risks to content, but whose assessment categories specifically take into account things like policies, governance and expertise, i.e. taking into account significant organizational and human factors [Dollar and Ashley 2013]. The role of maturity models is also being actively explored in the research data management domain, both at organization and community levels [Crowston and Qin 2010; Lyon et al 2012].

A similar approach has been taken by the US National Digital Stewardship Alliance in developing the NDSA Levels of Digital Preservation, which are understood to be "a tiered set of recommendations on how organizations should begin to build or enhance their digital preservation activities" [Phillips et al 2013]. The NDSA Levels provide technical guidance on preserving digital content "at four progressive levels of sophistication across five different functional areas," which are:

- Storage and geographic location
- File fixity and data integrity
- Information security
- Metadata
- File formats

The NDSA Levels are deliberately focused on the technical aspects of digital preservation as the team wanted them "to focus on practices, not policies or workflows, in order to allow immediate implementation" [Phillips et al 2013]. As it turns out, the functional areas identified by the NDSA correspond quite well to the types of information required for assessment at collection level.

4. COLLECTION PROFILING

The development of collection profiles at the British Library has been broken down into a number of smaller steps. The initial tasks were to identify the British Library's high-level digital collection areas and to develop an initial template to capture the required information [Day et al 2014].

4.1 Identifying high-level collection types

An initial practical task was to identify and define what we understood to be the Library's high-level digital collections. There was no agreed list of digital collection types held by the Library. Those lists that did exist - e.g. those provided by the catalogue or website - often included, for reasons of practicality, content types at several different levels of granularity.

In order to arrive at a more consistent list of candidate collection types, it was decided to supplement the categories found in these ad hoc lists with others derived from the Library's digital asset register. It is important to recognize that we were not trying to produce a definitive taxonomy of all digital collection types held by the Library, but simply to be able to identify collections at a sufficient (and logical) level of granularity in order to get started on the development of content profiles. The high-level collection types eventually identified (Table 1) included some that were firmly based on resource type (e.g. sound, multimedia), others that were multi-faceted but based on particular content streams (e.g. web archives); and others that followed more traditional categorizations of library collections, updated for the digital era (e.g. journals, books).

Table 1. Initial High-Level Collection Types

Type	Collection
Newspapers / journals	Digitised newspapers
	Born digital newspapers
Books	NDLP eBooks
	Voluntary deposit
	Digitised printed books
	Turning the Pages content
Manuscripts / Archives	Digitised Manuscripts
	Digitised archives
	Personal digital archives
	Turning the Pages content
Music	Digitised Music Collections
	Sheet Music
Maps	Digital mapping supplied by Ordnance Survey (GIS)
	Digitised maps
Academic journals	NPLD eJournals
	Voluntary deposit e-Journals
	Subscription e-Journals
Theses	Digitised theses
Patents	Patent databases
Web archives	UK Web Archive
	NDLP Web domain harvests
Sound / multimedia	Archive sound recordings
	Sound Archive (e.g., field recordings)
	Digitised sound / video
Stamps	Digitised stamps
Photographs	Digitised photographs
Printed ephemera	Digitised ephemera

The process of developing a list of high-level collection types, however, did raise some interesting questions about the task we had set ourselves.

4.1.1 Born-digital vs digitized content

For example, it was initially tempting to categorize digitized content separately from 'born-digital,' as this is a familiar distinction made by those considering digital preservation [Daigle 2013]. However, part of the aim of the profiling work was to try to deal with content by type, regardless of provenance or format.

So, for example, the British Library's digital newspaper collections would potentially include:

- Digitised printed newspapers from the Library's own collections (e.g. the historical newspaper collections digitized in collaboration with Gale Cengage, typically comprising images with searchable OCR text)
- E-editions of printed newspapers, ingested directly from newspapers' publication workflows (e.g. as PDF)
- Web-based newspapers captured as Web archives (e.g. newspaper websites captured as part of the UK Web Domain; the originals are typically constantly evolving Web pages with significant amounts of embedded content (e.g. images, video, surveys, comments) and links)

Obviously, within the Library's ingest and processing workflows these would be represented by quite different content streams, but the profiling activity does at least give an initial opportunity to consider all digital news content as a single collection, even if it is decided later on that more than one kind of preservation intent can be identified. Similar considerations would apply to other kinds of content.

At a more fundamental level, however, it is increasingly difficult to distinguish between born-digital and digitized content. As others have pointed out, much digital content is often a combination of several different kinds of content type, some of which may be born-digital, others not [Friedlander 2002]. This is perhaps most noticeable with Websites, but is increasingly true of many other kinds of content, For Example, e-journal articles or e-books could be understood to be simply containers for multiple kinds of content, which might include images, video, sound, games, software or data. In Europe, at least, research papers reimagined as compound digital objects (combining at least text and data) are sometimes known as "enhanced publications" [Doorenbosch et al 2009]. Eventually, as predicted by Kircz [1998], it might also be possible to think of all research papers as modular aggregations of many other kinds of content, including bibliographic information, content, abstracts, references, index terms, tables, etc., all of which could potentially have a different representation.

All of this meant that we needed - at least to start with - to focus on content type regardless of its immediate provenance.

4.1.2 The 'tangibility' of collections

When developing the collection profile activity, we also had to understand what exactly we meant when we talked about "collections"? Collections are a deeply embedded concept in memory institutions, so quite a lot of intellectual effort has been made over the years into trying to understand what they are and how they relate to wider organizational contexts. Traditional concepts of collection in library and information science have tended to focus on three main things: tangibility (regardless of format), ownership and a perceived user community [Lee 2000]. What has changed in the digital era is that library collections can be built without the inherent need for tangibility (although even digital content has to be stored somewhere) or ownership.

Like most other research libraries, the British Library routinely provides access to digital content that is not under its own direct control. As mentioned before, its current content strategy states that outside of legal deposit, voluntary deposit and donation, "the Library will prefer to connect to content," unless it wishes "to hold and preserve the materials for the long term" [British Library 2013a]. In this new collection management environment, active choices need to be made about precisely which content needs to become part of the permanent collections (and is thus able to be preserved). It is intended that the collections profiling activity at the Library will support collections management decision making, not least by gaining insight from collection specialists and curators on the specific preservation requirements of different classes of content. It might also help to clarify which particular content needs to become part of the Library's permanent collections.

Table 2. Initial Profile Framework Structure

Summary	Content Type (from list).
	Brief Description.
	Location.
	Curators / collection owners.
	Interviews held.
	Legal Deposit status.
	Creation status.
	Accrual status.
	Number of digital objects (approximate).
Background	An introduction to the content type, providing background on the collection/s covered by the profile.
Acquisition	Identifying the main current acquisition routes for collection content.
Preservation Intent	Summary of points agreed by curators / content owners, identifying the main characteristics of collections that will need to be preserved.
Acquisition Format	Identifying the main formats currently being acquired (where collections are complex, this does not need to be exhaustive).
Issues	Highlighting any specific current challenges.
Profile Metadata	Information about the completed collection profile itself, e.g. identifying creators, dates, and status / version number.

5. Developing the draft profile framework

The framework for the profile itself was developed at the same time as the identification of high-level collection types. The sections in the initial draft profile framework (November 2013) section headings were either generic (collection overview, preservation intent, rights) or broadly followed the functions defined by the Reference Model for an Open Archival Information System (ingest, archival storage, preservation planning, access control). Following the review of some draft profiles, the framework has been further simplified to reduce the number of sections required and to focus the profile on the key information types required to support the capture of digital preservation requirements (Table 2).

The draft framework was first introduced to and discussed with curatorial and other colleagues in the Library. It was then used to help create a number of draft profiles, initially for content types covered by Non-Print Legal Deposit content streams (e-journals, e-books, UK Web-domain harvests), then by a few selected others (manuscripts and archives, news content, sound content).

The profile framework will evolve further as we gain more experience with using it. Eventually, however, the plan will be to develop some support materials (e.g. documentation, a set of sample interview questions) that will help with ensuring consistency of approach. It will also be important to review the profiling process following integration with other preservation planning activities being undertaken by the British Library (e.g. file format assessments, tool assessments and preservation watch).

It is highly likely that both collections and preservation intent will change over time. There will be a need to ensure that collection profiling is undertaken on a regular basis and that it becomes part of the Library's business-as-usual digital preservation activities.

6. DISCUSSION AND OUTLOOK

The British Library's collection profile activity is an attempt to use content reviewing to capture information about collections and preservation intent to help inform digital preservation planning. Work to date has included an attempt to identify the high-level digital collections in the Library and to define an initial profile framework. Work on developing the profiles is ongoing as we progress in an iterative fashion. It promises to be an interesting approach, linking curators understanding of digital collections with the planning processes required to support their digital preservation.

7. ACKNOWLEDGMENTS

The authors would like to thank British Library colleagues for their comments on the draft list of high-level collection types and the initial collection profile framework.

Ann MacDonald was Intern – Digital Preservation at the British Library from October 2013 to March 2014 and was a key part of the team developing collection profiles.

8. REFERENCES

APARSEN. 2013. The British Library. In *Storage solutions summary of inputs*. APARSEN Deliverable D23.1, 26-33. (March 2013). Retrieved August 30, 2014 from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D23_1-01-1_0.pdf

Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman. 2008. Plato: a service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* (Pittsburgh, PA, USA, June 16 – 20, 2008). ACM Press, New York, NY, 367-370. DOI:http://dx.doi.org/10.1145/1378889.1378954

Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber and Hans Hofman. 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* 10 (2009), 133-157. DOI:http://doi.org/10.1007/s00799-009-0057-1

Caroline Brazier. 2013. born.digital@british.library: the opportunities and challenges of implementing a digital collection development strategy. In *Proceedings of the IFLA World Library and Information Congress* (Singapore, August 17 - 23, 2013). Retrieved August 30, 2014 from http://library.ifla.org/222/1/198-brazier-en.pdf

British Library Act 1972. Her Majesty's Stationery Office, London (1972). Retrieved August 30, 2014 from http://www.legislation.gov.uk/ukpga/1972/54/contents

British Library. 2013a. *From stored knowledge to smart knowledge: the British Library's Content Strategy 2013-2015*. British Library, London. Retrieved August 30, 2014 from http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf

British Library. 2013b. *Digital Preservation Strategy, 2013-2016*. British Library, London. Retrieved August 30, 2014 from http://www.bl.uk/aboutus/stratpolprog/collectioncare/discovermore/digitalpreservation/strategy/BL_DigitalPreservationStrategy_2013-16-external.pdf

CCSDS 650.0-M-2. 2012. Reference Model for an Open Archival Information System (OAIS). CCSDS Recommended Practice (June 2012).

Center for Research Libraries. 2007. *Trustworthy Repositories Audit & Certification: criteria and checklist (TRAC)*, v 1.0 (February 2007). Retrieved August 30, 2014 from http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0

Sheila Corrall. 2011. The concept of collection development in the digital world. In Maggie Fieldhouse and Audrey Marshall (Eds.). *Collection development in the digital age*. Facet Publishing, London, 3-25.

Sheila Corrall and Angharad Roberts. 2012 Information resource development and "collection" in the digital age: conceptual frameworks and new definitions for the network world. In *Libraries in the Digital Age (LIDA) Proceedings*, Vol. 12 (Zadar, Croatia, June 18 – 22, 2012). Retrieved August 30, 2014 from http://ozk.unizd.hr/proceedings/index.php/lida/article/view/62/33

Kevin Crowston and Jian Qin. 2010. A Capability Maturity Model for scientific data management. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (Pittsburgh, PA, USA, Oct. 22 – 27, 2010). ACM Press, New York, NY, Article 124.

Bradley J. Daigle. 2013. Dream the impossible dream: born digital stewardship. In *Archiving 2013 Final Program and Proceedings* (Washington, DC, USA, April 2-5 April, 2013). Society for Imaging Science and Technology, Springfield, VT, 2-5.

Angela Dappert and Adam Farquhar. 2009. Significance is in the eye of the beholder. In *Proceedings of the 13th European Conference on Digital Libraries* (Corfu, Greece, September 27 - October 2, 2009). Springer, Berlin, 297-308. DOI:http://dx.doi.org/10.1007/978-3-642-04346-8_29.

Michael Day, Ann MacDonald, Akiko Kimura and Maureen Pennock. 2014. Implementing digital preservation strategy:

developing content collection profiles at the British Library. In *Proceedings of Digital Libraries 2014* (London, UK, September 8 - 12, 2014). Forthcoming.

Charles M. Dollar and Lori J. Ashley. 2013. Assessing digital preservation capability using a maturity model process improvement approach. (February 2013). Retrieved August 30, 2014 from http://www.savingthedigitalworld.com/papers-research

Paul Doorenbosch, Eugène Dürr, Barbara Sierman, Jens Ludwig and Birgit Schmidt. 2009. Long-term preservation of enhanced publications. In Marjan Vernooy-Gerritsen (Ed.). *Enhanced publications: linking publications and research data in digital repositories*. Amsterdam University Press, Amsterdam, 157 -209. DOI:http://doi.org/10.5117/9789089641885

Luís Faria. 2013. Scout – a preservation watch system. In: *Open Planets Foundation blog*. (16 December 2013). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system

Luís Faria, Christoph Becker, Kresimir Duretec, Miguel Ferreira1 and José Carlos Ramalho. 2013. Supporting the preservation lifecycle in repositories. In *Proceedings of Open Repositories 2013* (Charlottetown, PEI, Canada, July 8 – 12, 2013). Retrieved August 30, 2014 from http://or2013.net/sites/or2013.net/files/PW_repositories_OR13_V0.5.pdf

Patrick Fleming. 2011. The British Library Newspaper Strategy: developing collaboration with publishers to digitise back runs and to ingest born digital newspapers. In Hartmut Walravens (Ed.). *Newspapers: legal deposit and research in the digital era*. IFLA Publications, Vol. 150. De Gruyter Saur, Munich, 21-30.

Any Friedlander. 2002. Summary of findings. In *Building a national strategy for digital preservation: issues in digital media archiving*. Council on Library and Information Resources, Library of Congress, Washington, DC, 1-8. Retrieved August 30, 2014 from http://www.clir.org/pubs/abstract//reports/pub106

Richard Gibby and Caroline Brazier. 2012. Observations on the development of non-print legal deposit in the UK. *Library Review* 61, 5 (2012), 362-377. DOI:http://doi.org/10.1108/00242531211280487.

Steve Hitchcock, David Tarrant, Les Carr, Hannes Kulovits and Andreas Rauber. 2010. Connecting preservation planning and Plato with digital repository interfaces. In *Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, 19 - 24 Sep 2010). Retrieved August 30, 2014 from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/tarrant-65.pdf

ISO 14721:2012. Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model. International Organization for Standardization, Geneva.

ISO 16363:2012. Space data and information transfer systems -- Audit and certification of trustworthy digital repositories. International Organization for Standardization, Geneva.

Jackson, A. 2012. Formats over time: exploring UK Web history. In *Proceedings of iPres 2012* (Toronto, October 1 - 5, 2012). University of Toronto, Faculty of Information, Toronto, Ontario, Canada, 155-158. Retrieved August 30, 2014 from https://ipres.ischool.utoronto.ca/proceedings

Joost G. Kircz. 1998. Modularity: the next form of scientific information representation? *Journal of Documentation* 54, 2 (1998), 210-235. DOI:http://doi.org/10.1108/EUM0000000007185

Gareth Knight and Maureen Pennock. 2009. Data without meaning: establishing the significant properties of digital research. *International Journal of Digital Curation* 4, 1 (2009), 159-174. DOI:http://dx.doi.org/10.2218/ijdc.v4i1.86

Hur-Li Lee. 2000. What is a collection? *Journal of the American Society for Information Science* 51, 12 (2000), 1106-1113. DOI:http://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1018>3.0.CO;2-T

Legal Deposit Libraries (Non-Print Works) Regulations 2013. S.I. 2013 No. 777. The Stationery Office, London (2013). Retrieved August 30, 2014 from http://www.legislation.gov.uk/uksi/2013/777/contents/made

Liz Lyon, Alex Ball, Monica Duke and Michael Day. 2012. Developing a Community Capability Model Framework for data-intensive research. In *Proceedings of iPres 2012* (Toronto, Ontario, Canada, October 1 - 5, 2012). University of Toronto, Faculty of Information, Toronto, Ontario, 9-16. Retrieved August 30, 2014 from https://ipres.ischool.utoronto.ca/proceedings

Peter May and Carl Wilson. 2014. *Technical architecture report*, v2. SCAPE Deliverable D2.3 (March 2014). Retrieved August 30, 2014 from http://www.scape-project.eu/deliverable/d2-3-technical-architecture-report-v2

Spencer McEwen and Andrea Goethals. 2009. File Information Tool Set (FITS): a new tool for digital preservation repositories. *D-Lib Magazine* 15, 9/10 (September/October 2009). Retrieved August 30, 2014 from http://www.dlib.org/dlib/september09/09inbrief.html

MIT Libraries. 2013. Digital Content Management Infrastructure Improvement: FY13 strategic objective. Retrieved August 30, 2014 from http://libguides.mit.edu/lifecycle

William Palmer. 2014. A Tika to ride: characterising web content with Nanite. In *Open Planets Foundation blog*. (21 March 2014). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/blogs/2014-03-21-tika-ride-characterising-web-content-nanite

Petar Petrov and Christoph Becker. 2012. Large-scale content profiling for preservation analysis. Poster presentation, iPres 2012 (Toronto, Ontario, Canada, October 1 - 5, 2012). Retrieved August 30, 2014 from http://ifs.tuwien.ac.at/~petrov/publications/c3po-poster-ipres12.pdf

Megan Phillips, Jefferson Bailey, Andrea Goethals and Trevor Owens. 2013. *The NDSA Levels of Digital Preservation: an explanation and uses*. National Digital Stewardship Alliance, Washington, DC. Retrieved August 30, 2014 from http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

Johan Van der Knijff and Carl Wilson. 2011. *Evaluation of characterisation tools, part 1, identification*. SCAPE Technical Report (September 2011). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011.pdf

Colin Webb, David Pearson and Paul Koerbin. 2013. 'Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collections. *D-Lib Magazine* 19.1/2 (Jan./Feb. 2012). DOI:http://dx.doi.org/10.1045/january2013-webb

Peter Webster. 2013. Crawling the UK web domain. In *UK Web Archive blog*. (16 September 2013). Retrieved August 30, 2014 from: http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html

Michael Witt, Jacob Carlson, D. Scott Brandt and Melissa H. Cragin. 2009. Constructing Data Curation Profiles. *International Journal of Digital Curation* 4, 3 (2009), 93-103. DOI:http://doi.org/10.2218/ijdc.v4i3.117

Sarah J. Wright, Wendy A. Kozlowski, Dianne Dietrich, Huda J. Khan, Gail S. Steinhart and Leslie McIntosh. 2013. Using Data Curation Profiles to design the Datastar dataset registry. *D-Lib Magazine* 19, 7/8 9 (July/August 2013). DOI:http://dx.doi.org/10.1045/july2013-wright