LIBER QUARTERLY 20 (1), September 2010 – ISSN: 1435-5205. P53–65
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

A Digital Library Feasibility Study

C. Henshaw, M. Savage-Jones, D. Thompson

Wellcome Library, 183 Euston Road, London NW1 2BE, UK,
c.henshaw@wellcome.ac.uk

Abstract

This paper presents the outcomes of a Digital Library Feasibility Study at the Wellcome Library. In particular, the study looked at the interoperability and integration between systems, including a back-end digital asset management (DAM) system with attached storage, a front-end delivery system, the use of METS to manage delivery of content, a full-text database with search engine, and a workflow management system.

Key Words: digital library; information systems; digitisation; indexing; METS; digital asset management (DAM)

1. Introduction

Over the next five years the Wellcome Library plans to transform itself into a ground-breaking digital library, with rich, dynamic content being made available around a series of strategic themes. The first of these themes is 'modern genetics and its foundations'. Digitisation of content relevant to this theme has now begun.[1]

The Library's digitisation strategy is highly ambitious, not only in its scale — to digitise over 30 million pages in five years – but also in its vision for how these items will be accessed and displayed. Specifically, the aim is to create a single repository which will hold images, full-text material, archives, videos, audio files, and born-digital archival materials (an 'integrated library'). This content will be presented to users via a rich and engaging interface that embraces Web 2.0 functionality.

The Wellcome Library currently does not have the infrastructure required to create an integrated digital library providing a seamless interface across catalogues and digital collections. In order to move forward in planning and preparing the system requirements for the digital library, a Feasibility Study was carried out between November 2009 and May 2010 to look at the key systems in more detail, and determine a way forward toward building this infrastructure.

2. Aims of the Feasibility Study

The primary aim of the Feasibility Study was to determine whether the library's existing digital asset management system — Safety Deposit Box (SDB), purchased to manage and preserve born-digital content — was an option for the management of digitised content on a large scale, including complex objects such as books and video. The Library also wished to test the use of JPEG2000 image files[2] in this infrastructure.

Secondary aims of the Feasibility Study included investigating the options of searching for and displaying that content. Questions around on-the-fly format conversion from JPEG 2000 for dissemination, the use of a cache for dissemination manifestations, the use of METS, and options for storing and searching full text were another focus of this study.

A final piece of work for the Feasibility Study was to investigate the possibility of procuring a workflow tracking system to manage the end-to-end workflow of all digitisation and born-digital projects. The Library came to recognise that large-scale digitisation, the ingest of multiple formats from multiple sources and the aggregation of descriptive and administrative metadata requires a level of management that would be unsustainable using current ad hoc collection-specific tracking and management mechanisms.

3. Methodology

The Feasibility Study was managed by the Wellcome Library's Digital Services Department, and work was commissioned from Tessella, the suppliers of Safety Deposit Box, and CCS Content Conversion Specialists GmbH (CCS), suppliers with experience in METS, JPEG 2000, full-text indexing,

and delivery of digitised content online. The following sections provide a brief summary of the methodology and key deliverables of the study.

3.1 Safety Deposit Box

As the Library was keen to keep its options open regarding the use of existing systems (as well as investigating new ones), a key element of this study was to determine whether SDB could be used as the back-end repository for managing files (in addition to managing the born-digital materials). In determining this it was necessary to investigate where SDB sits in the digitisation workflow, and what, if any, software modifications to SDB would be required to ingest and manage the digitised content.

Secondly, the Library needed to model and demonstrate the technical feasibility of accessing content held in SDB by a third party front-end delivery system.

Tessella was commissioned to undertake modelling and testing work using SDB version 4 (currently in late development stage) to answer these questions. The deliverables included:

- Final Report
  - Recommend modifications to SDB
  - Explain the protocols for importing and exporting content and metadata in SDB
  - Provide benchmark results for expected performance criteria including elapsed time
  - Provide hardware requirements to support the recommended implementation of SDB
  - Show how SDB fits within the digital library infrastructure on the whole
- Proof-of-concept system demonstrating the capabilities of SDB to ingest and manage digitised content and to make that content accessible to a third party system.

3.2 Delivery, Full Text and METS

To meet the goals of the study, the Library needed sufficient evidence to help it understand how digital objects held by SDB would be handled by the

delivery layer and the indexing layer, and how METS would be used to manage the structure and display of content. CCS was commissioned to provide information and recommendations on these elements, including demonstrating access to content held in SDB.

The deliverables of this piece of work included:

- Final Report
  - Recommend a METS implementation
  - Recommend a full-text index implementation
  - Recommend a front-end delivery system implementation, including on-the-fly conversion from JPEG 2000
  - Provide information on the interoperability work with SDB
  - Provide hardware requirements to support the recommended implementations
  - Show how these systems fit within the digital library infrastructure on the whole
- Proof-of-concept (using the Veridian digital delivery system) demonstrating the ability to request and receive content from SDB.

3.3 Workflow System

It became apparent during the digitisation planning stages that a workflow tracking system was required to manage the process of digitisation, track items undergoing digitisation, and to aggregate and output metadata as METS. The system will also track the ingest of born-digital materials. During the Feasibility Study the Library modelled a potential system and looked at a range of commercial products. The deliverable for this work was a report on how the workflow system would work in a practical sense, how it could fit within the digital library architecture, and a draft of the requirements that would ensure any proposed solution would meet the needs of the Library.

4. Outcomes

4.1 Safety Deposit Box

The SDB prototype successfully ingested JPEG 2000 images files, and a range of video formats using a mocked-up SIP (submission information package).

The SIP included a 'protocol' file — an XML file indicating to SDB that a set of objects was ready to be ingested from a specific location — and the content that comprised a 'logical object' (such as all the page images of a book), thus automating the process of ingestion into SDB.

Using JHOVE, SDB was able to characterise the JPEG 2000 files and add this administrative metadata to its metadata store for format preservation purposes. Characterisation of the Library's chosen video formats (mpeg, Windows Media Video and Quicktime) is currently not supported by SDB, although in future a new characterisation tool such as MediaInfo could be 'wrapped' and incorporated if required.

Using the Veridian front-end system, CCS was able to demonstrate remote request of files from Tessella using a proof-of-concept system, which, although mocked up, was realistic in the provision of files to a third party. The method used for the proof-of-concept was as follows:

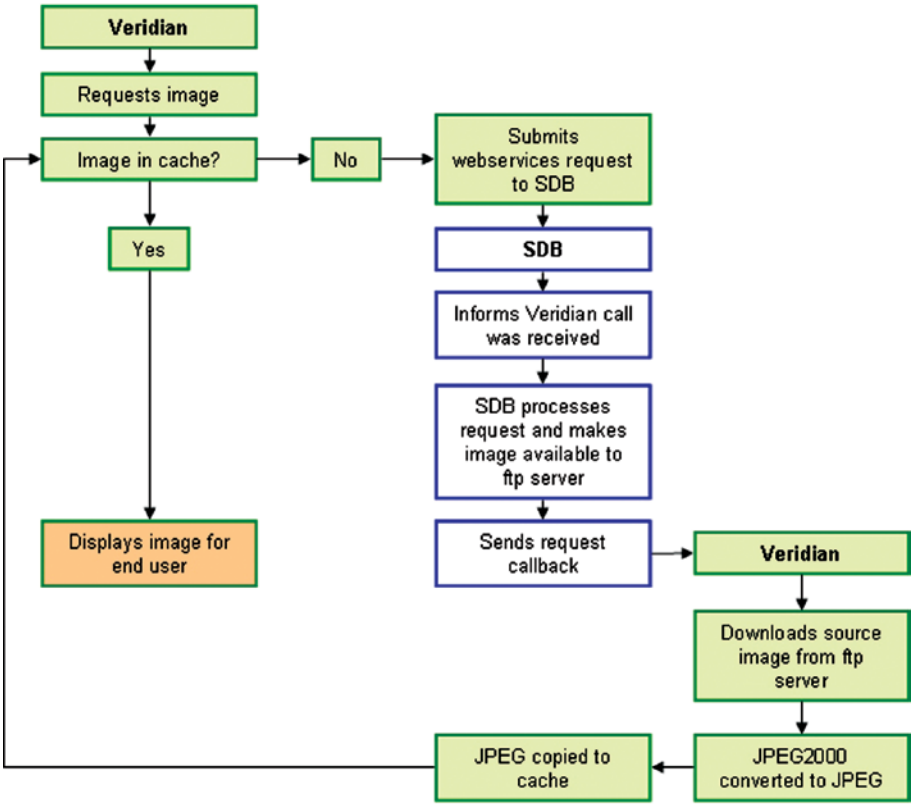
a) Veridian sends a submitRequest SOAP message to SDB requesting a DIP (dissemination information package) for a specific identifier, indicating, among other things, the name of the folder within the ftp site where the exported file(s) should be placed. The identifier could be either a logical object identifier, or the identifier for an individual file.
b) SDB acknowledges receipt of the request, packages up the files, and places them in the specified location.
c) SDB sends a JobCompleteRequest SOAP message to Veridian indicating that the DIP was exported.

A schematic of the proof-of-concept system infrastructure can be seen in Figure 1.

Further work:

- Investigate further Tessella's recommended modifications to SDB, including API's and ingest workflows.
- Compare the value of customising SDB to other potential DAM systems on the market.
- Carry out a full tender for the final system as appropriate.

Fig. 1: Proof-of-concept infrastructure used to demonstrate request of image files from SDB by a third-party system.



4.2 Delivery System

Neither the Library nor SDB has an existing digital delivery system that could demonstrate the features the Library wanted to model. There is a wide range of delivery systems on the market, but for the purposes of the Feasibility Study, CCS's Veridian system (developed by DL Consulting) was used as an exemplar to demonstrate both the requesting of digital content from SDB and as a working model for the infrastructure required for searching and displaying content, including full-text indexing and conversion of JPEG 2000s on-the-fly.

As described above, a proof-of-concept instance of Veridian was used to request images from SDB, acting as the proof-of-concept DAM. Once this content was retrieved from Tessella's servers, Veridian was able to convert the JPEG 2000 files into JPEG dissemination formats on-the-fly, which were stored in a cache. These images could then be viewed by a web browser, ready for zooming, panning, page turning and so on.

The Library intends to use a limited cache for dissemination files. When a digital file, or logical object is first requested by a user, the delivery system will have to request the JPEG 2000 from the DAM, then convert it to JPEG (or another format as required) and place the dissemination format in the cache. These would remain available in the cache for subsequent users, whose access would be quicker because the object had previously been requested and conversion already done. The cache can also be pre-populated or manually emptied if required. Veridian has demonstrated that on-the-fly conversion and use of a limited cache is feasible in practice.

The study did not look at website design, navigation, viewing/downloading options or Web 2.0 features, but focussed on the systems infrastructure.

Further work:

- Actual speed of content delivery to end users to be tested further using the Wellcome Trust's server and storage infrastructure.
- Complete specifications for the delivery front-end to be drawn up.
- Carry out full tender for the final system.

4.3 Full-text Index

The Library intends to OCR all printed materials that are digitised, and make the text searchable. CCS was tasked with providing recommendations for an indexing architecture that would meet the needs of the Library and which is to provide a quick, accurate full-text search for millions of page images.

CCS, using Veridian as an exemplar, recommended first of all the content that should be indexed. This included not only OCR but also ICR (Intelligent Structure Recognition), the use of dictionaries and word lists for tagging important keywords, and creating transcriptions or translations where possible to maximise the accuracy and relevancy of full-text searching. For

different types of content, the Library should set up profiles for indexing, to ensure that the methods used are appropriate to the text being digitised.

The indexing overhead would not be large for the Library's materials, which will usually consist of books of 250/300 pages, with no more than 500 words on a page. At a typical speed for indexing, this means around 150 books could be indexed per hour.

At its most basic level the indexing system consists of a database (the index) and a search engine. Solr is commonly used for full-text systems, based on a Lucene index providing fast results for large volumes of data, and for a large demand from users. For very large demands on the index, multiple servers can be employed.

The option exists to transcribe digitised archival collections before indexing them, or to include born-digital archival materials in the indexing process. This will be considered in a later phase of the project.

Further work:

- Investigate further the recommended indexing options with regards to improving search results for the collections to be digitised.
- Investigate further how the index could be accessed by the Library's current federated search system (Encore).
- Complete specifications for the full-text index to be drawn up.
- Carry out full tender for the final system.

4.4 METS

The Library anticipates using METS to facilitate the delivery of digital content online. The METS file will include an aggregation of metadata such as selected descriptive metadata from the catalogues, administrative metadata from SDB, and metadata recording access permission. Descriptive and administrative metadata held in Library systems will remain the authoritative source of metadata for resource discovery and material management.

During the Feasibility Study, the Library modelled a METS file according to the anticipated needs of the delivery system. A METS file will be created for each logical object, drawing together all the digitised elements that make up

that logical object. This could be pages of a book, segments and formats of a video title, or all the letters in an archive file, for example.

CCS provided further information on the use of METS and how the METS files could be structured. The key recommendation was that the Library creates its own METS profile. The METS profile provides a reference to the METS files the Library intends to use.

CCS also suggested that METS/ALTO should be used to provide structure to the OCR'ed text for printed items, in particular the ICR metadata as well as the word coordinates. METS and the ALTO extension are both used extensively in large-scale digitisation projects.[3]

CCS also recommended that the descriptive metadata in the METS file was held in Metadata Object Description Schema (MODS), while administrative image metadata was held as Metadata for Images in XML Schema (MIX).

Further work:

- Determine what metadata standard(s) to use for descriptive and technical metadata.
- Consider further the use of the ALTO extension for the content to be digitised.
- Finalise model for Wellcome METS profile, taking into account the needs of the chosen Delivery System, and commission the creation of the profile.

4.5 Workflow System

During the Feasibility Study, staff from the Digital Services Department modelled a 'workflow system' that would improve the management of the Library's intended digitisation program, as well as the ingest of born-digital materials. Originally, this was modelled on the National Library of Wales bespoke system, which not only tracked project activities (such as digitisation, QA, etc.) but also acted as an aggregator of metadata, producing METS files for use in displaying the content online.

As the study progressed, considering the metadata the Library needed to aggregate, and after looking at other examples of workflow systems, it

became clear that expecting a tracking system to also aggregate and output metadata as METS was quite a tall order. Most commercial systems would require extensive development work to make this feasible; something the Library wishes to avoid.

By the end of the study, the 'Workflow system' model was split into two potential systems: the 'Workflow tracking system' (WTS) and 'The Normaliser'. The WTS will be focused on the tracking aspects, and some limited data input according to the workflow diagram shown in Figure 2.

The WTS will carry out the following basic functions:

- Allow projects to be managed on a 'project', 'batch' and 'unit' level
- Associate descriptive metadata with each unit using barcodes
- Provide a graphical interface for input by project staff
- Authenticate users and user groups for different levels of access
- Track activities according to user input (a tick box for 'conservation completed', etc.) including dates/times such actions were taken and by whom
- Track the location of units according to user input
- Perform command line actions (such as converting images to JPEG 2000) where possible
- Allow for flexibility in workflow steps for different workstreams
- Store metadata in a standards-based database.

The Normaliser would exist as a separate application, independent of the other library systems, but utilising the WTS database. It would map incoming metadata from the Library's cataloguing systems (which use MARC21 and ISAD(G)) and the DAM into common database fields native to the WTS, thereby aggregating the metadata. The Normaliser would also map selected fields from within the WTS database to the DAM schema (to input catalogue metadata into the DAM for administrative purposes) and the Wellcome METS profile, producing output files. See Figure 2, above, which shows where in the workflow these input and output tasks occur. We anticipate that the WTS would run The Normaliser as a command line action.

A number of questions still remain, including the availability of existing applications that can provide this mapping function, whether library staff should be able to add and edit input and output XML mappings, and whether this

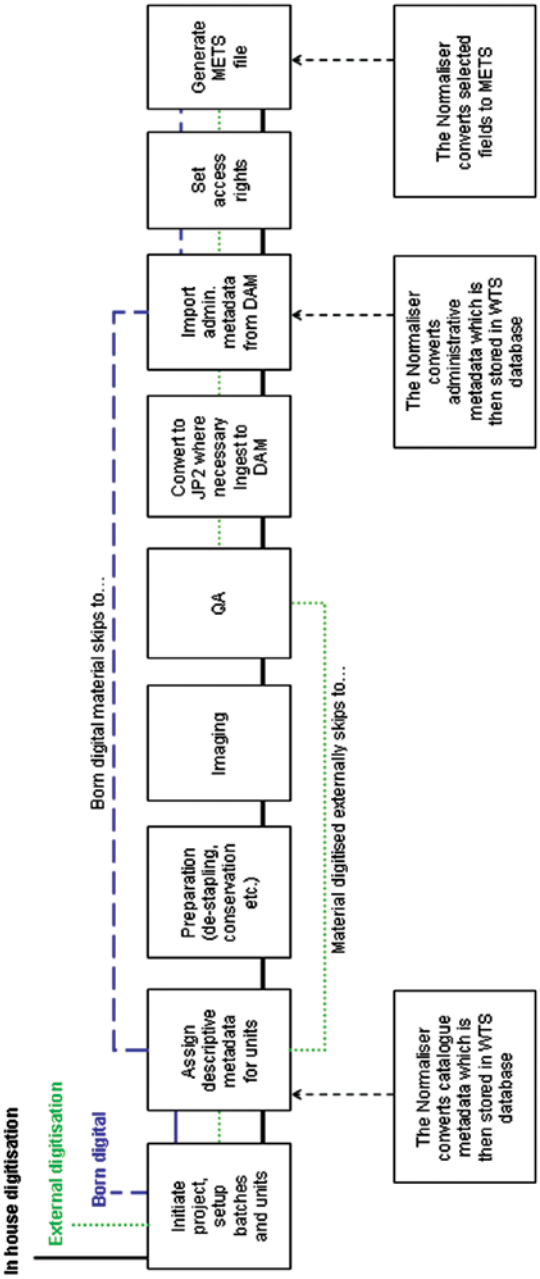


Fig. 2: End-to-end processes to be tracked for three work streams.

application should be solely a command line action, or include a graphical interface outside the WTS.

Further work:

- Complete the specifications for the WTS and Normaliser, taking expert advice on board.
- Investigate options for off-the-shelf products and bespoke systems.
- Carry out a full tender for the final systems.

5. Concluding Remarks

By carrying out a detailed Feasibility Study on its existing systems and the requirements for new systems, the Wellcome Library was able to form a plan to develop its digital library infrastructure over the next two years. Having taken the time to model, specify and test the interoperability and essential functions of these systems the Library can now move forward into the tendering stages with confidence that the major gaps and dependencies have been identified and addressed. Carrying out an end-to-end proof-of-concept, in particular, was extremely useful in highlighting these dependencies. Working with third parties for the majority of the Feasibility Study provided an object lesson in anticipating assumptions and communication needs when dealing with multiple suppliers on systems integration projects.

Websites Referred to in the Text

Alto, http://www.loc.gov/standards/alto/techcenter/use-with-mets.php

Encore, http://encoreforlibraries.com/

JHOVE, JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/

Lucene, http://lucene.apache.org/java/docs/

METS, Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/METSOverview.v2.html; METS profile, http://www.loc.gov/standards/mets/mets-profiles.html

MIX, http://www.loc.gov/standards/mix/

MODS, http://www.loc.gov/standards/mods/

National Library of Wales, http://www.llgc.org.uk/

SDB, Safety Deposit Box, http://www.digital-preservation.com/

Solr, http://lucene.apache.org/solr/

Veridian, http://www.ccs-digital.info/en/products/veridian

Notes

[1] http://library.wellcome.ac.uk/doc_WTX057852.html

[2] See http://library.wellcome.ac.uk/assets/wtx056572.pdf

[3] Mass digitisation projects in Europe, http://massdigitization.com/#europe. Also see the National Library of New Zealand's *Papers Past*, http://paperspast.natlib.govt.nz/cgi-bin/paperspast