DOI: 10.2352/issn.2168-3204.2016.1.0.127

Digitisation at Scale – Automating the mass acquisition of digitised content

Dave Thompson, Wellcome Library, London, UK

Abstract

This paper will set out to illustrate how the Wellcome Library[1] has created an end-to-end workflow that manages the acquisition of digitised content from numerous sources through the use of workflow tracking middle ware, a virtualized IT environment, providing for the storage of master images in a repository and final public access via the web. It will briefly talk to the use of a cloud based solution for storage and dissemination. It will show how acquisition can only work by the careful control of all stages of the process and by the application of high levels of automation. This paper will talk specifically about the use of management and tracking middle ware and to the design of processes and the relationships necessary to make the mass acquisition of digital content successful. The presentation will set out the processes of acquisition but also how to build on principles of digital preservation and life cycle management. The presentation will show that managing the high volume of the acquisition of digitised content is a matter of controlling each step of the process and learning 'on-the-job'.

Brief abstract

This presentation will set out to illustrate how the Wellcome Library has created an end-to-end workflow that manages the acquisition of digitised content from numerous sources through the use of managing and tracking software through storage in a repository to public access via the web. It will briefly talk to the use of a cloud based solution to dissemination. It will show how the process can only work by the careful control of all stages and by the application of high levels of automation.

Introduction

1. ***Motivation:*** The Wellcome Library is one of the UKs significant collections in the history of medicine. The Library is undertaking a long term digitisation strategy that will provide free public access to large amounts of content on-line. This strategy will transform the Library and its collections into a significant on-line resource. Not all of that content comes from within the collections held by the Library. The Library has formed partnerships with other library's and archives as well as with the Internet Archive[2] who are undertaking a significant part of the digitisation process. The ultimate goal being a cloud based system for dissemination of content through the Digital Library Cloud Service (DLCS) that the Library has established.

2. ***Problem:*** The Library is looking to create an on-line resource within the next 10 years. Whilst digitization may continue as a long term activity the aim is to create a large scale resource within a 'short' space of time through a coherent strategic program designed to maximize the efficiencies of working at scale. However, the Library has not previously worked at the scale required to build such a large on-line collection. The Library has previously worked with digitization only at a low volume. The anticipated creation/acquisition of digitized content was planned to scale to approximately 1 million JPEG2000 (JP2) images per month and the whole activity covering a number of years. Such a scale of activity required the use of new/additional software, adaptions to existing software and the design of new systems and processes to manage, store and disseminate the proposed volume of content.

3. ***Approach:*** The Library was clear that, wherever possible, existing systems would be used for the digitisation program. The Library is a long time user of the Preservica[3] digital object repository for born digital archival material; it made sense that the same repository was used for digitised images. It was clear though that new software tools, such as Goobi[3], would be required.

The initial approach taken was to set out a strategic program that encompassed the broad aims and a timeframe. It was this strategy that was accepted by senior managers and on which the program was based. It provided for a basis on which the necessary resources could be anticipated and sought. This in turn provided a framework within which a long term program could be designed.

The practical approach to problems of systems and architecture was iterative and based on learning as we progressed. Processes were initially designed based on the then current knowledge and on the systems and architectures as then specified. As these systems and processes have been used and as the project team has learned more, so the architecture and design of processes has evolved. Not evolved piecemeal but in a structured way in response to specific issues, opportunities or ambitions. A downside to this approach is that it might not always be possible to anticipate issues that are not directly visible or to implement 'solutions' to problems that are short term

DOI: 10.2352/issn.2168-3204.2016.1.0.127

or that do not address wider issues. Working at high volume doesn't necessarily imply that lots of different systems need to be used. In many ways the process requires that everything be as simple as possible and that there be as few 'moving parts' as possible; the level of simplicity achieved is of course relative. This also relies and can be built on close relationships with system vendors who were required to provide not only system support but also system development. It also required vendors to 'buy into' the strategic vision and to help design workflows and tools that specifically support working at high volumes without becoming overly complex or reliant upon an interaction with large numbers of humans.

4. Implementation

To support the key aim the Library uses Preservica as its digital object repository application. This repository is used to store 'master' images which ultimately will be made available through the Library's Digital Library Cloud Service (DLCS). It was quickly realized that manual processes would not achieve the levels if ingest that were being planned.

The key to achieving high volume throughput is a piece of middle ware called Goobi[4]. This is work-flow tracking and management software. Goobi is capable of processing large volumes of digitised content but not at the levels the Library planned for. It was not possible to simply employ additional staff to increase throughput.

The approach taken was twofold; firstly to automate processes as much as possible and, secondly, to design and implement IT infrastructure to maximize throughput. The desire for automation combined changes to system architecture with process redesign. When looking at processes – especially ones involving humans – the role of users was carefully evaluated and where possible previously human tasks were automated or workflows redesigned to reduce the need for human involvement. For instance, the process by which digitised content was FTP'd to the Library was integrated with Goobi so that the ingest process was automatically triggered by the arrival of content in a watched folder on the FTP server.

Implementing better process management relied on expanding the Goobi Intranda Task Manager (ITM) as both load balancer and plugin manager. The highest volume of digitised content is material automatically harvested from the Internet Archive[4] (IA) website. A package of four files is automatically downloaded by Goobi, comprising image files, descriptive metadata, structural metadata and raw optical character recognition (OCR) files. Having these files, this specific data set, means that the process of acquisition, processing and storage can be fully automated.

When performing high volume processing Goobi's typical installation placed a lot of the processing burden on the application server, e.g. in creating JPEG images or in processing the raw OCR files to create ALTO files. To spread this load additional servers were introduced on which the resource heavy sub-systems were installed. The Goobi architecture allows for individual workflow tasks, such as image conversion, to be run as services on these separate sub-systems. Sub-tasks were implemented as plugins on the ITM and this way their management was centralized into a single management tool. The simple UI to this tool made it easy to manage individual processes and to shut down individual processes without affecting the rest. This has made daily maintenance simpler by allowing specific services to be shut down when, for instance Wellcome Trust IT need to perform maintenance.

Equally it was important that when processing content at volume it was very difficult to do quality assurance on anything but a small percentage of the total volume being processed. Goobi has a number of validation steps within its workflows that check that numbers of images being processed match the expected numbers of derivatives being created. For instance the number of ALTO files must match the number of page images. If any process fails this validation then the workflow is automatically halted and a human can check on what the problem is. On top of this the Library implemented its own validation step to ensure that the JP2 images being processed were valid ad properly formed. To do this software called the Jpylyzer[5] was installed. Building this validation step into each workflow meant that every JP2 being processed by that workflow was tested to ensure that it was valid and well formed. Adding this step of automated quality assurance has 'captured' instances where incorrect file formats have been submitted or where files have failed to be created to our specifications.

5. Results: The results are clear and content increasingly publicly available on the Library website. It is possible to manage the acquisition of high volumes of digitised content from a diverse range of sources by applying careful process and system design whilst at the same time applying high levels of automation. For the calendar year 2015 an average of 1.1m images per month were processed by Goobi.

What has emerged has been a relatively complex IT infrastructure comprising a number of individual servers that support the core application Goobi. However, this complexity represents the simplest approach that we can currently devise. The iterative approach of developing an initial process, testing that in production use and revising in the light of experience has proved itself. Our approach has worked well as the Library had little experience to start with. It is a flexible approach and one that can

quickly respond to changes and/or new developments; e.g. to changes in the IA website. The ability to adapt IT infrastructure has also been key. Working in a virtualized IT environment it is easy to create a new virtual server and, more importantly, to give any virtual server the resources necessary for a specific task. The use of a virtualized environment also meant that the Wellcome Trust IT department could perform routine maintenance e.g. server patching, more easily by being able to 'move' virtualized servers between hosts. Equally it is easy to take down a virtual server if that approach has not shown benefit and little capital expenditure is lost. This ability to 'experiment' is a valuable tool in when looking at process design.

It was also key to be working with a vendor who was flexible in their approach to systems support and who was prepared to make changes to their system on an on-going basis. In this sense there was learning for both the Goobi vendor and the Library. There was success in another way. By managing processes and by achieving high levels of automation a high level of consistency was achieved in the way data was being processed. This meant that images processed were consistent in format, validity and the means of their processing. From a life cycle management perspective this has produced a coherent body of material that can be more easily managed in the future because of its consistency.

As the volume of content being processed increases the need for storage of the master images also increases. Preservica is the repository software. To date this has used local storage within the Wellcome Trust. However, as content increases it becomes more difficult to justify the use and cost of local storage. In early 2106 the Library started a project to store digitised content in the cloud.

6. ***Developing approaches:*** It is not always easy to do process and system design in an iterative way. But equally it can be frustrating to have to stick to plans that were developed before a project started and which can be shown to benefit from being changed. Experience has taught us that our plans at the beginning - plans based on theoretical assumptions about working with high volumes of content were (mostly) correct.

   It has been crucial that we have a good working relationship with our Goobi system vendor Intranda. Without their understanding and support it would have proved very difficult to take an iterative learning approach to process design. Intranda provides the Library with the Goobi software, with support for that software and development services. We have relied on Intranda to apply their technical knowledge and experience of working with their customer base to help us. As a piece of middle ware Goobi has proved to be flexible and adaptable. However, what has been more important has been Intranda's willingness to work with us to both identify issues and to implement solutions. It can be a slow process to identify exactly what an issue is and then to devise a remedy. In some cases the changes we made were wrong and created more problems than they solved and we had to back track. This can be frustrating both to the Library and to the vendor who is being asked to make changes. However, we have been able to design better processes and systems after practical experience in their use. In turn this supports Intranda in that they gain wider experience in the use of their software, experience that they can offer other institutions and other Goobi users.

   The approach taken has resulted in processes that are specifically tailored to particular acquisition activities, e.g. Automated harvesting from the IA website. This has proved to be the most complex activity. The content we seek resides on the IA website. After work to identify what material needed to be harvested the process design looked at how data was structured on the IA website and how these might be harvested. The approach was iterative but had to be specific to the types of material; single monographs are different to multi volume monographs and both are different to journals. The process design looked at what workflow steps to use and in what order they should be used. It looked at creating scripts that would take the data from the IA and automatically process it. In this case the automated IA workflow built on experience gained from simpler processes and adapted it.

   The automated harvesting of multi-volume works has proved the most difficult process to design because of needing to maintain the relationships between each individual volume. This piece of process design is on-going.

   In some cases harvesting material from other institutions meant that different problems were encountered. These might be the use of different identifiers within the data package that we were seeking to download, missing files, or entirely missing items. In cases of problems human intervention was required to identify the issue and to manually apply a remedy. The benefits of high levels of automation are suddenly clear when having to manually resolve problems and at times our 'snagging' lists have grown quite large. Work on this particular workflow is continuing. To date it has not been possible to develop a more generic or universal model of automated harvesting from remote sources. However, we have iteratively improved process efficiency and throughput. We have eliminated mouse clicks, unnecessary steps and applied more and more automation as we aim to acquire more content and as we learn to trust our systems.

7. ***Conclusions:*** The Wellcome Library is not yet half way through the period envisaged by the initial strategic plan. Progress has been good but not without

DOI: 10.2352/issn.2168-3204.2016.1.0.127

some issues. However, our initial assumptions about working at scale have (Mostly) stood the real world test. The Library has a high degree of control over the means by which digitised content is captured, processed, and stored. We have a high degree of control over Goobi and its sub-systems and a high degree of flexibility over the infrastructure we have created. A unified data set has been created in which JP2 images are uniform and should prove more easily managed when that format becomes obsolete. Already careful system and process design is delivering results. The mass acquisition of digitised content can be achieved but only within the framework of an overarching strategic plan that has been agreed to at the highest level within the organization. This strategy was then implemented with high levels of control over the processes involved and increasing amounts of automation used in those processes. Just as important was the flexibility and adaptability of working within a virtualized IT environment. The ability to spin up additional servers when required, to give them resources for specific task and to do so in an economical way has made managing the IT infrastructure simpler which has allowed a focus on the proper work of running a digitisation program.

The strategic plan not only provides a basis for resourcing the activity but allows a degree of flexibility in that the agreed strategy can be applied in an evolving and flexible way. The strategy provided overarching measures of success but didn't prevent the activity from becoming more efficient though an evolving process based on experience and practical learning. This organizational buy in and commitment at the highest level has ensured that resources were available when they were required.

References

1. Wellcome Library. http://wellcomelibrary.org/ Accessed 3 February 2016
2. Internet Archive. https://archive.org/ Accessed 3 February 2016
3. Preservica. http://preservica.com/ Accessed 3 February 2016
4. Goobi. https://www.goobi.org/en/ Accessed 3 February 2016
5. Jpylyzer. http://jpylyzer.openpreservation.org/ Accessed 3 February 2016

Author Biography

Dave Thompson is Digital Curator at the Wellcome Library in London. He has worked in information management for the New Zealand government, as an independent contractor in libraries in both New Zealand & the UK. Dave began looking at data management in about 2000. Dave holds a BA (Hons) in history from the University of Wales, St David's University College & a MA in Library & Information Studies from Victoria University of Wellington, New Zealand.