Managing the Digitisation Chain: Practical Tools for Process Management and Multi-Faceted QA

Nele Gabriëls; KU Leuven Libraries, ULS Digitisation & Document Delivery; Leuven, Belgium.
Dirk Kinnaes; KU Leuven Libraries LIBIS; Leuven, Belgium.
Diederik Lanoye; KU Leuven Libraries, ULS Metadata Services and Acquisitions; Leuven, Belgium.
Bruno Vandermeulen; KU Leuven Libraries, ULS Digitisation & Document Delivery; Leuven, Belgium.
Mark Verbrugge; KU Leuven Libraries, ULS Digitisation & Document Delivery; Leuven, Belgium.

Abstract

The complexity of the digitisation process in a decentralised environment requires profound workflow and risk management, with attention for solid communication and for QA across the digitisation chain. In the absence of an all-in solution, ongoing projects at KU Leuven Libraries have served as a virtual lab environment to gradually create and test specific tools and standardise workflows since 2009. The resulting set of step-specific QA tools and processes form an integral part of a standardised yet modular digitisation workflow. The modular setup provides flexibility when developing project-specific workflows. Detailed tracking of the complete digitisation process for each individual project through a workflow management system allows for shared communication as well as for overall high-level risk management and multi-faceted QA in a time-efficient manner.

The Challenge of a Decentralised Environment

Digitisation is a complex process encompassing many aspects and involving various actors. The recognised core activities of imaging, upload and preservation indispensably require at the least physical preparation and transport, metadata creation specifically for digitisation, and publishing to achieve the impact that digitisation strives for. In KU Leuven Libraries' working context, all digitisation activities are performed in a decentralised organisational structure: people, collections and infrastructure are scattered across several locations and departments. This significantly increases the importance of workflow and risk management. As a medium-sized library with existing library information systems deeply embedded in its processes, an all-in solution tracking e.g. transport, metadata creation, data creation, file movement and integrity, and upload, proved unfeasible. Manual workflow management, on the other hand, is both time-intensive and insufficient for risk management. Progress reports may not be filled in, transport tracking is diffuse, communication remains outside the tracking environment and based on individual's thoroughness and good will.

What is needed are tools for heightened quality assurance and risk management as well as a way to manage workflows and communication more efficiently. The widely recognised core activities of transforming physical objects into digital data and of data upload and preservation are guided by international standards such as Marc21/Dublin Core, Fadgi, OAIS and Premis, enhanced with various validation tools, controlled workflows and automated processes. Literature covering QA in these areas is ample and for that reason out of the scope of this paper. However, QA is crucial not only in these but in all aspects of the digitisation process, where it may not only concern technical aspects. Having started in 2009 with project-based digitisation, KU Leuven Libraries gradually designed its digitisation process and accompanying QA through a process of trial and error. One development would always induce the next. What follows describes the evolution of the in-house, photography-based digitisation of documentary heritage at KU Leuven Libraries.

File naming and metadata

KU Leuven Libraries embarked upon the digitisation journey with just one staff member concerned with digitisation of the physical collection in general, a partnership with the Faculty of Arts' photographer, and a collaboration with the library information system (LIS) provider LIBIS holding the ExLibris preservation environment DigiTool in its portfolio. With access to both imaging equipment and infrastructure for storage and preservation available, the digitisation process during these first years started with what seemed the most logical step: photographing books. Image capture, processing and file renaming was followed by metadata creation in the metadata department and upload, publication and preservation by the LIS provider.

When the collection items to be digitised increased both in numbers and complexity (and as a result an imaging assistant joined the Lab), checking the accuracy of the imaging and metadata, and correcting errors became increasingly time-consuming. Common mistakes such as missing pages, incorrect file names, incomplete metadata or wrongly ordered images often were discovered when books had already left the Imaging Lab or even after ingest. In addition, the imaging operator was required to make decisions as to what exactly needed capturing without having the required expertise to do so concerning the historical and physical characteristics of early prints and manuscripts. The question of whether or not (and how) to capture stubs, loosely inserted leaves, foldouts, moveable parts on the page (e.g. volvelle), etc. slowed down the process and was not always answered successfully nor in a consistent way.

As a first step, meaningful file names with shelfmarks and image description (front cover, folio numbers, …) were replaced by unique numeric file names: 'digi-numbers'. Digi-numbers included the project number followed by an object number (e.g. DIGI_0004_0012). For complex objects with many images, a sequential number was added (e.g. DIGI_0004_0012_000097). Ever since, files could be ordered through correct file naming. Any potential issues caused by incorrect page sequences in the original (e.g. when a fascicle C would mistakenly follow a fascicle D) or by digitised collection items changing (physical) location could also be eliminated.
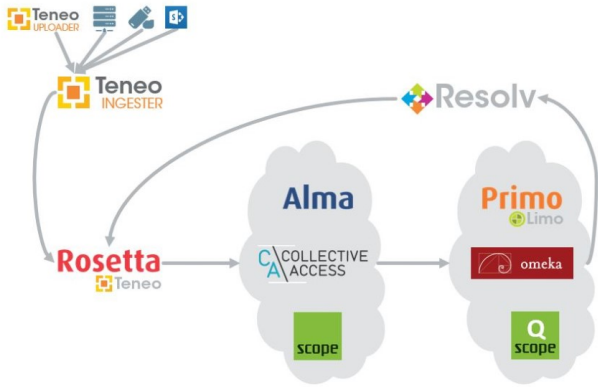
Figure 1: Current preservation and publication systems' architecture.

Shortly after introducing a numeric file name for the images, the metadata creation was moved forward in the digitisation process. Metadata for digitisation consisted of a list of concordances that enabled the connection between the descriptive metadata of a single object in the library's metadata repository (at first Aleph, later Alma) and the corresponding digital object in the preservation environment (at first DigiTool, later Rosetta) through the Digi-number. It also included a structural map per object: a full list of individual file names and corresponding meaningful labels (e.g. folio numbers) to enhance the user's consultation experience.

With metadata creation preceding imaging, complete lists of objects to be digitised and structural maps for each object defining all individual images were being drawn up, including expert and uniform decisions and precise guidelines for capture in the Imaging Lab. Imaging operators now received detailed instructions leading to a significantly higher level of efficiency, accuracy and consistency within the Lab. By following the list of images during capture, they in turn also verified the accuracy of the structural metadata.

Digicorder for metadata QA

Figure 2: Semi-automatic metadata creation in Digicorder.

In order to heighten efficiency and reduce the risk of human error when drafting the metadata, a tool for system-controlled metadata creation was developed in-house: Digicorder. This FileMaker-based application reduced keying to a minimum. It allows for semi-automatic creation of structural metadata, e.g. generating series of filenames and

labels that are uniform across all projects, even if these labels are complex codicological collation formulas. Additional comments for the imaging staff may be added to the system by metadata experts to complete the capture information.

Digicorder proved also important for two other reasons. First, following capture and processing, the correspondence between structural metadata and images requires final verification before upload. Missing image files or metadata entries and incorrect combinations of images and metadata are easily detected thanks to Digicorder's preview facilities for visual QA, through an HTML file containing the structural metadata entered by the cataloguer as well as links to the images on a network drive where it harvests thumbnails. The result is an easy interface for metadata experts with a side by side visualisation of images and structural metadata.
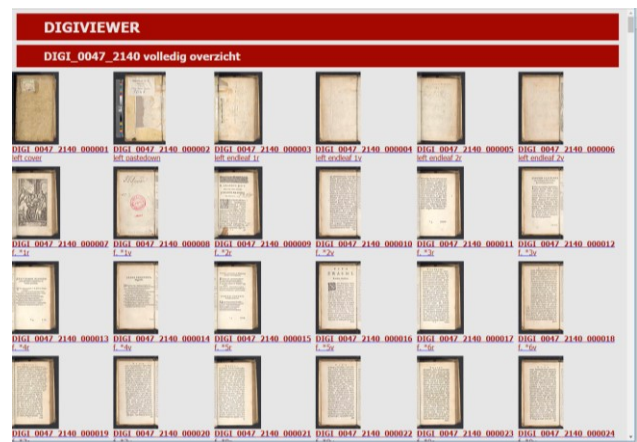

Figure 3: Digiviewer for visual QA of combined metadata and imaging output.

Second, Digicorder's controlled environment not only ensures a high level of uniformity and accuracy of the structural metadata throughout the digitisation activities. It also automatically exports this metadata into the correct file template and format and even positions it in the appropriate directory for data delivery towards ingest, when the Digicorder metadata is combined with the images to form the digital objects that are uploaded into the digital repository.

Workflow standardisation

After a few years, the number of volumes going through the digitisation process increased significantly thanks to digitisation becoming part of the library's general policy. Tools and standards for QA of four core areas (metadata creation, imaging, upload, and preservation) were by then in place and continuously evolving. With the rising numbers and the resulting gradual expansion of library collections and people involved, several aspects needed clarification and fine-tuning. The set-up for each new project was time-intensive, as digitisation was not yet standardised into a well-defined process.

Based on the default process outline that had organically grown from past activities, all steps were scrutinised and structured. The resulting standard workflow is composed of seven components: preparation, metadata, imaging, enrichment, ingest, publication, and post-care. [1] It has a modular set-up: each of the components may be moved and adjusted depending on the project specifications and the demands of the materials.

For graphic materials (with one object being 2 images at most) imaging comes before metadata, for example.

Each of the components clusters a number of actions that are assigned to a specified digitisation role (metadata provider, imaging operator, ingest operator, publication executor, collection curator, project manager). The preparation phase deals with project application and planning, legal agreements, external item loans and insurances, transport issues, and physical pre-/conservation actions. Existing descriptive metadata is analysed, completed and corrected during the metadata phase, and structural metadata for digitisation is created. Imaging encompasses both capture and processing, while ingest includes merging metadata and images into digital objects and their upload. The publishing phase links/exports the digital objects with/to the chosen publication platform. Digital preservation is the core of the post-care phase, together with archiving any documentation (e.g. overview of objects digitised). Enrichment (e.g. OCR or advanced imaging techniques) may or may not be added to the workflow. Its position and content depends entirely on the digitisation question.

The standard workflow can be applied to most of our documentary heritage digitisation projects. Since its introduction, it has been serving as a production line to which physical items can be fed so as to produce corresponding digital objects that are added to the digital repository and published through the library's public catalogue or other platforms. Project set-up is significantly simplified and introducing newcomers to digitisation became easier thanks to the standard procedures in place for each of the components.
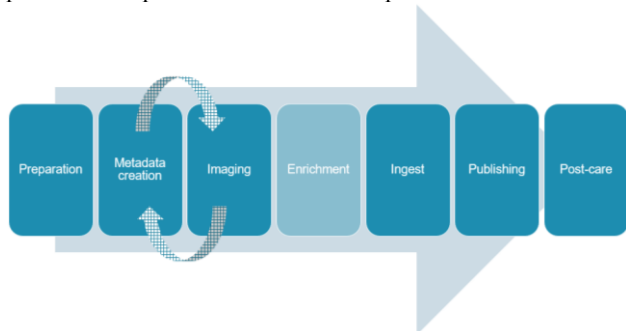

Figure 4: Modular digitisation workflow.

In order to control the increasing numbers, physical objects are no longer treated separately but gathered into packages during metadata creation. The size of these packages is defined at the start of a project and depends mainly on the number of images that can be produced in the Imaging Lab in a specified and manageable time-span (often one week, amounting to 3000-5000 images). Digicorder counts the total number of requested images and indicates when packages are full and ready to be transported to the Imaging Lab for imaging. Packages remain together both as physical and as digital objects throughout the digitisation process. That way, the risk of losing track of a single physical or digital object, already diminished by the concordances lists produced through Digicorder, is further reduced.

Process and communication management

Standardisation proved to be of major significance for risk management. In defining the workflow, the transitions from one component to the next were exposed as the weakest links. The progress of items through the digitisation chain was being tracked manually in project spreadsheets where everyone involved in digitisation indicated the status of his/her work for each separate collection item. There were two major issues with this procedure. The first was the lack of up-to-date information in the spreadsheet due to forgetfulness, the second concerned how each of the digitisation collaborators would know when to initiate their allotted actions. One depended on receiving a private phone call or email from the person performing the previous action. A trustworthy process overview proved impossible.

In order to strengthen the digitisation chain, it was crucial for transition points and information flow to be secured. Searching for a workflow management system (WMS), both specialised all-in systems for digitisation process management and general, off-the-shelf software were taken into consideration. As an example of the first, Goobi was considered.[2] The integration of operational and tracking functions was most promising but the software integration required for its implementation was not feasible due to the limited resources available and because of the impact it would have on the LIS and tools already in place.

Three off-the-shelf WMS's were tested in early 2015: Smartsheet, MS Project and Teamwork Projects.[3] All three were either web-based or could function on existing infrastructure and met the minimal requirements of low cost, automatic email notifications and the possibility to include subtasks. The decision to implement Teamwork projects was based on the availability of features such as project data export and project reporting, on the intuitive interface and on excellent user support. Important, too, was the full in-house management and independency from KU Leuven's existing systems, so as to be able to add users without administrative and other issues.

The standard workflow has now been set up in the WMS. First, for each digitisation project the various project roles (e.g. imaging operator, metadata creator, ingester, collection curator, project coordinator) are assigned to project collaborators. A package template is created that reflects the digitisation chain with a standard sequence of tasks and subtasks. Not every individual action is translated into a single (sub)task. Within the WMS, the main focus is on the unstable transition points that follow a sequence of tasks performed by a single project role. Each of these points is made explicit and traceable in the WMS. Some of the transition points correspond to the end of the process components (e.g. when the metadata, imaging, and ingest phases are completed), others are in fact the release of data/actions for review by others (e.g. visual QA of joint metadata and imaging output by the metadata creator, after capture; manual spot checks during the publication phase by the collection curator). Each task in the WMS is assigned to one of the project collaborators and task progress followers are defined. This ensures email notifications are automatically sent to the appropriate persons. Task specifications (e.g. details on file directories, file naming, task targets, etc.) are included in the task description for easy reference.

Based on this template, a separate task list is created for each package. Every member of the digitisation team has access to the full project in order to follow its progress and registers the completion of his/her assigned (sequence of) actions by ticking the box. The workflow is now largely self-directed: intervention is reduced nearly entirely to issue handling. As alerts and problem descriptions are recorded through the WMS's comments feature and automatically sent to the

appropriate collaborators, these issues have now become traceable, allowing for efficient process management. With all communication centralised within and directed through the WMS, the transition points linking the elements of the digitisation chain have stabilised.
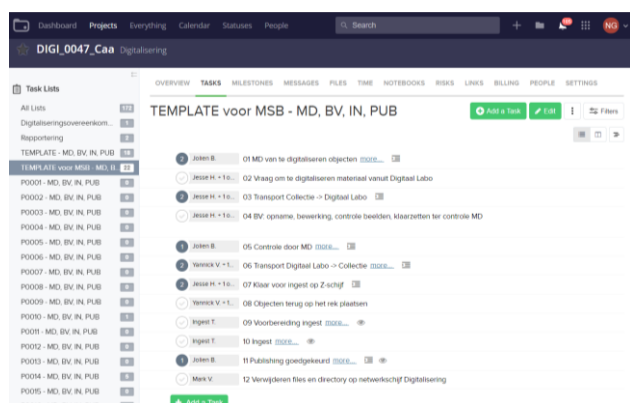

Figure 5: Project management through WMS: package template for project DIGI_0047: Collectio Academica Antiqua.[4]

A never ending story

Improving the digitisation process is a never ending story. Defining the workflow has brought to light other weak links than the ones described above. They are dealt with depending on impact, urgency or opportunity. Tracking physical materials, for example, remains a point of attention. The WMS includes a task to notify receipt of the materials both in the Imaging Lab and, after imaging, at the collection holder. Due to recent experiences, however, a paper registry was reintroduced in the Imaging Lab to record incoming and outgoing collection items, especially for scan on demand request (fast in-out).

Another example is tracking the physical integrity of collection items during the digitisation process. During the metadata phase, all collection items are subjected to a basic physical object evaluation. The Imaging Lab in addition has extensive experience with manuscript digitisation (both high-quality standard digitisation and advanced imaging techniques) but to date the physical condition of these documents has not been explicit monitored throughout the process as advised in Korthagen 2016.[5] It is one of the focuses of a pilot project for the integration of conservation and digitisation activities when dealing with fragile manuscript materials.

Conclusion

Throughout its digitisation activities, KU Leuven Libraries has continuously developed tools and processes to assure consistent and high-quality digitisation output. In order to achieve this, QA must be part of all digitisation component as well as of the overall process. Simply designing a standardised workflow

identified weak links in the chain for which new tools and approaches were sought. The implementation of a WMS proved to be a crucial instrument for communication, with all partners now permanently up-to-date with both workflow progress and issue sharing. Key aspects such as metadata creation, imaging, standard publication and preservation each have their own tools for QA but are now also firmly embedded in the workflow tracking. Problem registration and solving too are recorded in the WMS. Overall, the result is a flexible, modular system for versatile QA and process management, allowing for a standardised yet tailor-made process development depending on the digitisation project's objectives. Thanks to these implementations and developments, time spent on process management and issue correction has strongly been reduced. Process management through the WMS is now being applied to all digitisation efforts at KU Leuven, including the scan on demand services or bespoke projects for external clients. New digitisation processes and issues raised in existing flows remain permanently evaluated for future additions or corrections to the standard workflows.

References

[1] https://enrichingheritage.wordpress.com/2016/09/22/large-scale-high-end-digitisation-workflow-optimalisation/.

[2] https://www.intranda.com/en/digiverso/goobi/goobi-overview/

[2] https://www.smartsheet.com; https://products.office.com/en-us/project/project-and-portfolio-management-software; https://www.teamwork.com/project-management-software.

[3] https://expo.bib.kuleuven.be/exhibits/show/lovaniensia.

[4] Ilse Korthagen et al., Digitalisering van manuscripten. Checklist (Metamorfoze / Universiteit van Amsterdam / KU Leuven, 2016).

Author Biography

Nele Gabriëls received her Ph.D. in Musicology from KU Leuven (2010) and holds a post-graduate degree in Library & Information Science from Antwerp University. As Domain Responsible Digitisation at KU Leuven Libraries, she works on policy development and process improvement.

Dirk Kinnaes is Master in Engineering: Computer Sciences. He is a digital archiving and preservation specialist and Senior Project Manager at KU Leuven Libraries, LIBIS.

Diederik Lanoye holds a MA degree in History from KU Leuven. He is Process Manager Metadata Services and cataloguer of early printed materials at KU Leuven Libraries.

Bruno Vandermeulen holds an MA in Fine Arts, Photography from the LUCA School of Arts, Brussels. He is an imaging specialist in the field of cultural heritage and archaeology, focusing on art-technical imaging, and Head of Digitisation and Document Delivery at KU Leuven Libraries.

Mark Verbrugge is Master in Modern History. He is Reference Librarian in the field of science, engineering and technology and Digitisation Project Coordinator at KU Leuven Libraries with specialist knowledge in intellectual property rights.