DOI: 10.2352/issn.2168-3204.2016.1.0.107

Quality Assurance in Mass Digitization Projects

Martina Hoffmann; National Library; The Hague; The Netherlands

Abstract

QA can vary from simple procedures to highly developed workflows. Within one of the largest digitization programs in the Netherlands (Metamorfoze) the National Library has the task to ensure the quality of digitized images for preservation. To accomplish that task the National Library has taken several steps and is constantly improving its own process of quality management to ensure high speed, high volume and high standard controls for a huge amount of terabytes of data that will be stored permanently and made available for (online) use. In this paper you can find the questions we had to answer in order to set up the QA workflow, the workflow we did implement, the current status of our workflow and the lessons we learned along the way.

Metamorfoze – 2D mass digitization of cultural heritage

Metamorfoze is the Netherlands' national program for the preservation of paper heritage. The program started in 1997. The Metamorfoze program is hosted by the Koninklijke Bibliotheek (National Library of the Netherlands). Within the program there are two large sections based on different originals namely: the printed works which include books, magazines and newspapers and the more diffuse collections which include archival material and all other (museum-) paper collections.

Both types of originals differ in appearance as well as in desired output (access copy). What they have in common is that the digital masters have to replace the originals in daily use and the originals have to be stored responsibly in suitable storage materials to preserve them. The digital masters are the starting point for access copies which are mostly made available online. However, there are a lot of differences in the demands for each section: Where OCR is a standard requirement for printed books and newspapers, it does not add much for e.g. a manuscript from 1600. This is why we need different approaches to quality management for the two sections. Where we demand a lot of (standardized) metadata for a digitized book, we hardly obtain metadata from the digitization parties on the contents of each image in the archival section. In the archival section the content is provided by the collection owner, and covers the whole collection or a large section within a collection and therefore lots of images. Also the role of National Library differs a lot in both sections. For the printed materials the National Library is in charge of the whole process of digitization as well as the digital preservation. The digital output is managed and stored in the storage facilities which are owned by the National Library, and the Library makes the access copies available online. In the archival section of Metamorfoze, the task of digital preservation is carried out by the National Archives and the collection owner has to ensure that the images are made available in a suitable way. The various collections in the archival section are made available either online or offline on-site. Here, the National Library just has the task to ensure the quality of the preservation masters and therefore only carries out the quality assurance workflow, while the collection owners remain responsible for the whole project as project leaders. Within the quality assurance process, the senior production manager digitization of the archival section is responsible for implementing, improving and carrying out the quality control workflow. She also has to ensure that the public can find and use the output generated in the course of those digitization projects.

The archival section - Why do we do it?

Unique fragile material such as manuscripts and handwritten journals, letters and drawings are the originals that are digitized within the projects. The objects originate between approximately 900 to 1950. Some came from the former colonies of the Netherlands and were shipped several times in the past. The physical state due to poor storage conditions or excessive handling requires conservation before digitization can start. As mentioned above, digitization in Metamorfoze projects has to replace the original object in daily use. This is precisely why we demand the highest standards from the digital images. After being properly conserved and digitized, the originals are retained, but stored in dedicated facilities, and are no longer be available to the public.

How can we control quality in an efficient way?

We can produce millions of images at high speed, but without a certain level of quality control, we run the risk of wasting serious amounts of money. Therefore, it is important to take several steps before starting a large digitization project. First comes thinking about the final goals. That means we have to answer these questions: what do we want to create, and for what platform/use/public? This is actually a good start for any project and it is a necessary step before setting up any quality control workflow at all. Quality management is a process within a wider context, and controls the consistency of the desired output and helps maintain the speed in mass digitization at the lowest cost. Therefore, the first steps required are deciding what the goals are and what routing will be taken towards those goals. As case study for this paper we take the Metamorfoze Archival section, but the steps towards an efficient workflow for quality control are applicable to any (mass) digitization project whether it is 2D or 3D, high quality-high speed, high quality-low speed, access driven or for long term preservation and many more. We have to think what the desired output is for our images and determine how we can achieve this result. The first question you should ask yourself is whether or not there is a need for high quality-high speed digitization based on input (original material) and output (final digital goals). If your objects are not fragile, in good shape and can be digitized over and over again there is probably no need to produce masters for long term preservation at the highest measureable best practice. Maybe ‘readable’ is a good standard for your goals. Based on the output there are lots of questions to ask – and each and every one of those will influence the decisions you have to make in your quality control process. For example:

- What kind of material do we have?
- Which guidelines are available? In 2D mass digitization there are a few guidelines that can help maintain the

DOI: 10.2352/issn.2168-3204.2016.1.0.107

quality of an image such as Preservation Imaging Metamorfoze [1] or FADGI [2] but there is no comparable consensus on metadata, post processing and other areas which are important for a good quality image.
- One can ask if there is a necessity to use the Preservation Imaging Guidelines Metamorfoze? Maybe FADGI is a better choice or maybe the Guidelines for photographic materials [3]?

We need to know what we want to do with our images:
- Will they be stored in a long term preservation storage only or do we want to put them online?
- Do we only need an access copy?
- What are the needs for metadata and in what level? When we store only for long term preservation one can choose not to embed all kinds of metadata with the collection in the storage but to use any kind of database instead. Also the amount of metadata will probably differ according to the original object.

As for production itself:
- Can we do it in-house or do we need external suppliers?

It is obvious that the control process will differ depending on the decisions made in that stage. The control process can also differ in extend and amount of time needed to ensure quality standards are met based on the desired output. Therefore it is necessary that the aims are set before thinking about the quality control.
For the archival section of Metamorfoze the quality control process is built for the following requirements:
- Highest possible standard imaging (preservation imaging) due to the fragility of the objects we digitize.
- Ensure each image is suitable for long term digital preservation.
- Reliable connection between originals, user copy and long term master images.
- Processing high volume of data in a short time.
- Reliable and communicable results for all parties.
- External production line, digital storage and project management.

How do we do it?

A good quality assurance process incorporates all of the necessary areas and operates semi-automated, because computers can take over a large part of the workload.
In our archival section we work with different cultural heritage institutions from the Netherlands with different materials and with different digitization parties. We have standardized a workflow for digitization of unique material and the quality control which is situated in the National Library. The National Archive is the party for long term preservation. In this case we have to deal with requests from three different parties in the quality control workflow to ensure the final goals are met as is stated above.
The National Library uses a model which consists of three basic steps in her quality control workflow. However, there is a fourth step to be executed by the cultural heritage institutions themselves.

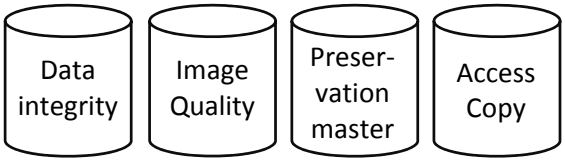

Figure 1. Four step QA process for Metamorfoze archival digitization projects

Step 1: 'Data integrity' is to check the data for the usability of long term preservation and metadata such as headers or the relation between originals, masters and access copies.
Step 2: 'Image quality' is to check the image quality and settings of the hard- and software that is used by the different digitization parties and to ensure that the guidelines are met. In our case the Guideline Preservation Imaging Metamorfoze.
Step 3: 'Preservation master' is a check on the output images for long term preservation storage. Within this step there is also a check for the different demands of a cultural heritage institution.
Step 4: 'Access Copy' is the step where the National Library delivers the derivatives from the preservation masters to the cultural heritage institutions to check the quality of the content and for their own usage. They get preservation master image quality images as a starting point for publications or to make their collection available on their website. On those derivatives they can do all the software enhancements that are necessary based on their needs. The National Library does not take part in this step and excludes these derivatives from the quality assurance process.

Step by step

We ask the digitization parties to help our quality control process by providing several extra features for each batch of images:
- Daily targets as required by the guideline Preservation Imaging
- Daily measurements of the targets
- Masters with object level targets as required by the guideline Preservation Imaging
- Masters without object level targets that are cropped and delivered as requested by the cultural heritage institution who is the owner of the digitized images.
- Derivatives for access as requested by the cultural heritage institution.

Furthermore, the National Library has written a specific guideline for 'submission of digitized materials in the Metamorfoze archival section to the National Library' where required file structures, requested header fields, file formats and many more items with regard to long term preservation and the relation between originals and preservation masters are defined and explained (Dutch only).

In order to be able to set up an efficient quality control process we had to establish first, what the key demands are with regard to the final result that we want to achieve. In our case the long term storage is the final result of the process. Therefore we put the check of the usability of the data at the start of our process: If we cannot use the data in the end we will reject the data at the beginning without further checks. To ensure this is established as quickly as possible, the National Library made an effort to automate this step as much as possible. Although there are several software packages on the market we have discovered that a lot of the applications are not suitable to our mass production process with hundreds of variables. With experiences from the past years

DOI: 10.2352/issn.2168-3204.2016.1.0.107

in mind, we have taken the step to optimize our process with software specially designed for our needs. We built modules based on existing scripts and software that can be switched on and off according to the input that we get from our projects and we can insert individual variables before starting the controls. We use GUI's that are easy in use and generate automatic reports from all controls. High volume digitization demands high speed quality control which is why we choose strong computers and do a lot of testing and benchmarking in software that is available or built to our needs. In the past year we were able to cut the computer processing time in half which allows us to get faster to step 2 in our quality control. We still try to optimize step 1 'Data integrity' further by thinking of more logic that we can script into software. However, this is only possible with the cooperation of our digitization parties who have to adjust their output according to our demands. For this to work, cultural heritage institutions also have to take their responsibilities by including a simple step in their preparation processes. As with all large projects for digitization it is important to know what we want to digitize – not only for an institution and possible funding but also for the external partners in such a project. The digitization parties and also the (external) quality control party has to know beforehand what they can expect and what data to expect. We therefore ask the cultural heritage institutions to fill in lists with the content of the collections they want to have digitized and checked by the National Library. We validate the lists before the digitization is started on technical issues such as unique inventory numbers or that only a limited set of numbers and characters is used to name objects.

Our quality assurance process is built upon those lists and we check all incoming data against them. With those lists we can track through al steps in the process.

Once we verified that the incoming data belongs to one of our projects we start the standard tests like reading out checksums or validating the filenames and formats. But we also built in a special test that allows us to match each image to its according daily target and vice versa. We follow the image quality standard Preservation Imaging Metamorfoze guidelines which have two quality levels for handwritten material: Metamorfoze light or Metamorfoze. We have to ensure that the image quality as described in these guidelines is met. It is therefore absolutely necessary that we can link the right targets to the images. What we also built is a standardized structure for each collection. According to the lists from the institutions our digitization parties can create a standard file tree which follows the levels of an archive. The unique identifier of an institution is the starting point of our file tree as well as each file name. In its most simple structure we have three levels – institution; collection; dossier where the third contains the actual image files with a sequential incrementing number at the end. Of course, more complex collections have more levels but we can fit most archival collections in this structure. The structure of the archive is also the structure of our file tree. To ensure correct delivery we have created software that will read out the hard drives for us and check all details that were specified before like the unique identifier for an institution. Where every original has its place in this standard structure, its corresponding digital image has its place in the file tree. The relation between input and output can therefore be established with a reliable connection and will be preserved in the digital storage of the National Archives.

The results of our computer processed control step provides us with a lot of information that we need in the quality assurance of the images themselves. Some examples: we do know already what color space our images will have and can adjust our calibrated monitors accordingly. We do know which targets relate to which images. We do know that no files are corrupt. Step 2 'Image quality' can take place. In this step 2 we check the daily targets from our digitization parties through own measurements (such as Slanted Edges, Digital Colorchecker SG, etc) and we crosscheck this with the delivered measurements from the digitization parties. As the guidelines state clearly which values each target should have we can control the settings of hardware and post processing of the images in an objective way. Nevertheless it is important to check images e.g. for artefacts or unwanted pixel disturbances caused by post processing. To ensure quality throughout a daily production the Guidelines Preservation Imaging Metamorfoze require for each image to be placed with an object level target which is also measured and checked by our staff.

After completing step two we can conclude that the data we received can be stored in the National Archives e-depot (a digital repository) as that the images taken meet the standard of high quality as required by the Preservation Imaging guideline of Metamorfoze. We can also identify each file submitted by its unique filename and properly relate it to its original which lays in a storage facility of the owning cultural heritage institution. What we don't know is whether the images are suitable for access and whether they have been delivered according to the wishes of the institution that will provide the access copies to the public. Typical publication means include websites, on-site computer facilities or portals such as APEx [4]. There is a big difference between the quality of an image in technical terms as hue, sampling efficiency, sharpness, color accuracy or geometry which are described for the preservation masters in the Guidelines Preservation Imaging Metamorfoze and the demands of an access copy where readability, landscape orientation, high contrast and many other factors are important. While readability is not a quality demand as such – for what are the specifications for readable? – we can conclude that with high image quality as Metamorfoze demands the captured texts will be readable. It does however not provide any answer to the orientation of an image. It may be delivered upside down or in a ninety degree deviation. For access output it is very important to know if the image has to be further processed (for example by rotation) or if spreads have to be split before uploading. Such requests are made for each project by the institution and can be different for every collection. Those decisions are made based on desired output and further handling by the institution itself. The National Library checks those requirements in a third step in her quality control 'Preservation Masters'. We take a sample according to DIN ISO 2859 part 1 and check each image of the sample in Adobe Photoshop on actual pixel size for orientation, cropping, overall appearance, artefacts – which can be missed in step 2, and the wishes as specified by the collection owner in an earlier stage of the project. If a batch is not passing all three steps in the National Library the batch will be rejected by the senior production manager, a report will be sent to all parties and the batch has to be repaired by the digitization party before entering step 1 of the quality controls again. If a batch of images passes all steps of the quality control process of the National Library a positive report will be sent to all parties and the data is transferred to the collection owner, who will take another step (step 4) and check the content of the images. That last step assures for example that the right image is placed with the right number in the digital storage facility of the National Archive. Further processing will however be needed for an institution to use the derivatives for publication or printing. Contrast enhancements or color scheme changes are not allowed for preservation masters

but an institution can carry out those adjustments according to the purpose they need the images for after completing the digitization project Metamorfoze.

The first three steps of this quality control process allow us to maintain high speed controls for each batch delivered by external parties. We define a batch as an external hard disk with a capacity of 2 to 3 TB data. On average a batch contains 1.8 TB of data that we process in approximately 12 hours plus a 24 hour circle of computer processing. It is possible to reject batches at any stage of the quality assurance process. Our reports are built up the same way as our process and each uncompleted stage in the report shows clearly that the step in control is not taken yet because of errors in the step(s) before.

What are the key lessons for a good quality assurance we learned?
To come as far as we are now in our quality assurance process we invested years of experience and lots of practice and whenever we thought we were on track we discovered another big step that we could make in efficiency. We still do! Our process is highly efficient as we processed about 400 TB of data, containing over 10 million digital images from approximately 50 different projects, many of them with their own specifications and difficulties in the past year. We expect that we can do twice as much with the latest efforts we put in renewing software and maximizing the speed of our controls. In the past years we learned that such a high speed process on high quality level depends on several factors within an organization.

- We learned that there is a huge need to use a common language. Therefore we are at the moment busy to expand the Guidelines submission of digitized materials in the Metamorfoze archival section with a glossary so that all parties in the whole digitization process can refer to the same vocabulary.
- We learned that a process needs high level of support from management. Quality assurance means, in the perception of many, extra time and extra costs for a project. Therefore it is important that management sees the value and supports the quality control. Setting up a QA process means thinking through all possibilities beforehand and create an environment of hardware, software and manpower that works as an efficient production line. Such an investment is only possible when there is a great support from the top as well as the bottom. Management is key to setting up an environment which is the basis of an efficient workflow. A good process manager or quality assurance manager needs support from the top of the organization to be able to set up the best possible workflow. Investments need to be made and improvements are required through time because the techniques and available options are rapidly changing.
- The starting point for any quality assurance process has to be a thorough evaluation of the different needs and wishes (stated and unstated) from all parties involved. More parties (internal and external) means more demands in checking and control throughout the process.
- It is necessary to stay informed and educated as quality assurance process manager about the newest developments on an international level about hardware, software, preservation issues, metadata and many more. If you can't stay on top of the development yourself - hire the expertise as needed.
- There is a lot of software available, but most software does not do exactly what we want and what we need. One has to optimize the existing software. In many cases, simple adjustments can make a huge different in processing time.
- We also learned that our process depends on knowledge not only of the process itself but also deeper knowledge of each step of the way. If you don't have the knowledge in your organization - hire experts.
- If you work with external partners for digitization in a project where a third party or fourth party provides the original input in any possible way you have to work together with your suppliers in order to create a steady workflow and keep the machines running.
- Look for smaller processes to improve ongoing rather than taking on the whole process at once.
- Be transparent and clear in your demands, standards and quality control.
- There is always space for improvement!

References

[1] Guideline Preservation Imaging Metamorfoze; https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf

[2] FAGDI Guidelines; http://www.digitizationguidelines.gov/guidelines/digitize-technical.html

[3] Guidelines Digitisation of photographic materials; http://en.nationaalarchief.nl/sites/default/files/docs/guidelines_digitisation_photographic_materials_0.pdf

[4] APEX – Archieves Portal Europe; https://www.archivesportaleurope.net/home

Author Biography

Martina Hoffmann is Senior Production Manager digitization at the National Library in the Netherlands for the archival section of Metamorfoze. She was operational manager quality control of digitized products in the National Archives in the Netherlands. She co-designed several quality assurance workflows for different mass digitization projects in the Netherlands. Starting with only image quality QA processes her main focus now are QA processes including several fields of expertise from metadata to long term preservation.

