ARCHIVING CONFERENCE
IS&T imaging.org

https://doi.org/10.2352/issn.2168-3204.2018.1.0.22

Dos and Don'ts for digitisation workflows

Steffen Hankiewicz; intranda GmbH; Göttingen, Germany

Abstract

Digitisation projects are generally complex and usually turn out to be more time-consuming than initially expected. The exact nature of the desired results should be determined well before the project starts – partly because whoever is funding the project will usually have made this a condition of support, but also because many similar projects will already have been carried out and these can be used as a guide. Yet many digitisation projects are launched without harnessing the available synergies. New software is implemented. Data formats are redesigned. In some cases, the entire system of project organisation is reinvented and tailored to meet the demands of a single project. Why?

This paper describes some of the typical pitfalls associated with digitisation project workflows and explains how even very large projects can be managed without reinventing the wheel.

Familiar territory

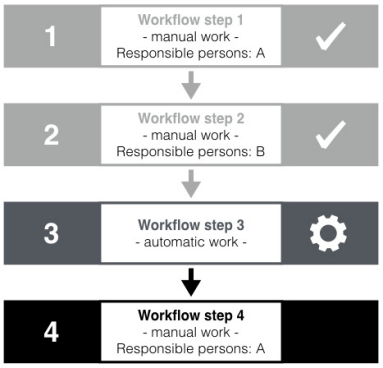
Digitisation has been a routine procedure in cultural institutions for over twenty years. With the possible exception of 3D objects, various approaches have become firmly established over the years, covering a huge range of objects. If we take the digitisation of books, for example, we can see that many approaches have caught on and are used today in the majority of digitisation projects. In most cases, objects are now digitised in colour, usually with a resolution of at least 300 dpi, and where possible stored in lossless formats such as TIF and JP2. The standardised ALTO format is widely used for text recognition output. In the ideal scenario, a full description of the digital product – together with all its derivatives, full texts and descriptive metadata – is provided in a standardised METS file.

At least that is the theory. Things have become a great deal more complicated since the days when digitisation projects merely involved producing scans or photographs. The process of taking high-resolution photographs and storing them in file systems is generally just the start. That's when the meticulous work of quality assurance and data enrichment begins. At the end of the day, the most frequent objective for most projects is to make the digitised material available to a wider public audience. On their own, images in computer directories are insufficient. That is why we need people with a range of skills who can work together to produce the desired results. In this context, it is important to consider that some tasks only make sense at a particular stage of the project. They may depend on a previous step and for this reason cannot be performed until later. It is precisely this need for coordination – who should carry out which tasks at which exact point and for which object? – that makes digitisation projects so complicated. Quite apart from the challenges thrown up by the object's physical properties and the need to produce valid digital data formats, it is above all the sheer volume of data and the correct sequencing of numerous small tasks that dictate the project's overall scale and complexity. This is precisely where we should enquire about lessons learned in other projects and examine how those projects were approached.

What are the most common approaches in use today?

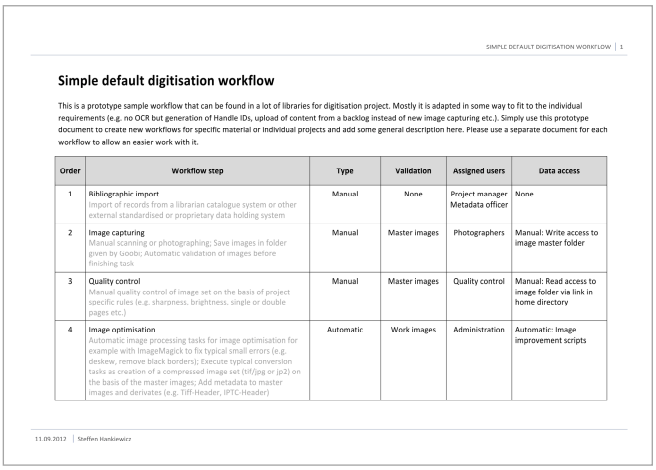
It is of course difficult to draw meaningful comparisons between large and small institutions, especially when they are based in different countries and have different budgets, skills and other resources. Nevertheless, it is larger digitisation projects that provide the most useful lessons. The processing of huge amounts of material in a series of workflows places great demands on projects in term of throughput, efficiency, automation and of course validation to systematically detect and avoid errors. This is also where the greatest organisational challenges arise – orchestrating the many tasks that make up the project as a whole. In smaller projects, however, it is often the demands of enriching digital output that prove more difficult to meet. The challenge here is one of detail, since the materials are often characterised by diversity and individuality rather than their sheer volume. Put simply, we can say that small-scale projects tend to focus more on depth, while large projects are more concerned with breadth, i.e. throughput. Despite these contrasting objectives, however, the approaches adopted by both large and small projects are essentially similar in that the workflow for each object (e.g. a book) is systematically divided into smaller individual steps, each of which has an associated set of responsibilities, a prescribed sequence and a mixture of both manual and automated tasks. Projects basically need a 'to do' list for each object that can be worked through line by line over the course of the workflow.

In general, as long as tasks and responsibilities are divided in this way and steps are taken to ensure that the results are properly documented (ideally in a central location), it is highly likely that the project will be implemented in a robust fashion since every step is traceable – provided also that the theoretical planning leading up to the project matches the day-to-day working routines of those involved.



Good planning is half the battle

Thorough up-front planning is crucial to the successful implementation of any digitisation project. At this stage, we can avoid the need for time-consuming changes to the workflow further down the line by setting out all the individual tasks that need to be performed for each object being digitised and defining precise responsibilities and potential validation routines. Experience shows that a simple workflow planning template in tabular form, ideally produced in consultation with the entire project team, provides the most suitable basis and avoids most potential misunderstandings further down the line.



SIMPLE DEFAULT DIGITISATION WORKFLOW | 1

Simple default digitisation workflow

This is a prototype sample workflow that can be found in a lot of libraries for digitisation project. Mostly it is adapted in some way to fit to the individual requirements (e.g. no OCR but generation of Handle IDs, upload of content from a backlog instead of new image capturing etc.). Simply use this prototype document to create new workflows for specific material or individual projects and add some general description here. Please use a separate document for each workflow to allow an easier work with it.

Order	Workflow step	Type	Validation	Assigned users	Data access
1	Riblingraphic import Import of records from a librarian catalogue system or other external standardised or proprietary data holding system	Manual	None	Project manager Metadata officer	None
2	Image capturing Manual scanning or photographing; Save images in folder given by Goobi; Automatic validation of images before finishing task	Manual	Master images	Photographers	Manual: Write access to image master folder
3	Quality control Manual quality control of image set on the basis of project specific rules (e.g. sharpness, brightness, single or double pages etc.)	Manual	Master images	Quality control	Manual: Read access to image folder via link in home directory
4	Image optimisation Automatic image processing tasks for image optimisation for example with ImageMagick to fix typical small errors (e.g. deskew, remove black borders); Execute typical conversion tasks as creation of a compressed image set (tif/jpg or jp2) on the basis of the master images; Add metadata to master images and derivates (e.g. Tiff-Header, IPTC-Header)	Automatic	Work images	Administration	Automatic: Image improvement scripts

11.09.2012 | Steffen Hankiewicz

At this point, it is also worth reviewing existing workflows that are being used for other current projects and setting them out in the same tabular form. The active involvement of the full team is vital to the project's eventual success, since this approach means that long-established methods are automatically reviewed at the same time.

Dos and Don'ts

Learning from other in-house or external projects with comparable requirements is tremendously beneficial, especially where the methods used by those projects have become established as best practice. At the same time, they are a source of valuable information that can be difficult to find in articles and books – information that is all the more effective if it is passed on through personal contact or during visits and conferences. The resulting advice and support can take many forms and cover many areas, e.g.

- software tools
- hardware equipment
- room design
- advice on object processing
- photographic/imaging techniques
- data formats
- interfaces
- workflow sequences
- advice on quality control.

Alongside such tips on best practice, observations of numerous successful, large-scale digitisation projects repeatedly highlight various key points – things that must be avoided at all costs and others that really do need to be carried out. Based on our experience of advising on digitisation workflows at over sixty cultural institutions in fifteen countries, we are currently preparing a set of guidelines (dos and don'ts) for publication. The guidelines will cover every aspect of digitisation and highlight suitable options. Some examples are given below.

Workflow and planning

Do remember that your output should be designed to last.

In most cases, the output of a digitization project should remain usable for a long time. Furthermore, since digitization projects often extend over a long period, all workflows should be organized in such a way that the plans can still be followed much later. Directory structures, filenames and formats should be as clear and simple as possible, ideally self-explanatory. Future project staff should always be able to locate documentation describing and explaining each step of the workflow to ensure a smooth transition in the event of personnel changes.

Do break down the digitization workflow into small and logical steps.

Workflow tasks that need to be performed for every object (e.g. for every book) should be broken down into smaller steps. Each step should be planned in such a way that it can ideally be completed relatively quickly by appropriately trained staff in the designated user groups. Rather than:

- "Digitize a book, extract and record the metadata and publish it online",

it would be better to break down the workflow as follows:

- Person A: Scan (book by book).
- Person B: Extract/record metadata (book by book)
- Person C: Publish (book by book)

Using this approach, the status of each object is clear and verifiable throughout the workflow. It also greatly increases productivity, as project staff repeatedly perform the same task for multiple objects without having to familiarize themselves with a different role.

Do harness the experience of project staff in your planning.

Although staff who have spent a great deal of time working on other projects may be reluctant to adopt new methods, their knowledge and experience is tremendously valuable when it comes to redesigning workflows or planning new ones. You will need to harness that experience so that you can design an effective new workflow that covers every single task from start to finish. It follows that the workflow should never be imposed "from above" by management without input from project staff. This participatory approach to workflow planning also promotes greater acceptance of new methods and tools.

Do plan in advance for platform-independent digitization workflows.

There are some very good software tools that can only run on certain operating systems (e.g. Windows or Mac). Ideally, you should plan your workflows from the start in such a way that individual steps can also be performed on other operating systems. This gives you maximum flexibility over your future working methods and avoids the risk of "vendor lock-in".

Do use a central workflow tool to coordinate your project.

Ideally, the tasks performed for each digitized object should be precisely verifiable so that you can maintain a clear overview of progress and any hold-ups and monitor any recurring errors and the circumstances in which they arise. Microsoft Excel and a central network drive with write access for each member of the project team are not usually adequate for this purpose. For this reason, you should consider whether to adopt a professional work-flow management tool. These are available under an open-source license and with a large community of users.

Do think about the intended results well in advance.

You should consider as early as possible what is to be done with your digital output. Even during the actual workflow, you can make preparations for subsequent online publication, transfer to another technical system, or ease of access for researchers. Validation, data conversion and metadata enrichment can all make a crucial difference in terms of the way the material is used.

Do ask your scanning service provider to make interim deliveries.

If you choose to outsource the job of scanning to an external provider, you should ideally request a series of interim deliveries rather than one large final delivery. This will help you to ensure that the outsourced work is consistently of the required quality and to keep check on your overall project schedule. Ideally the contractor should upload the data directly to your system rather than supplying a hard disk.

Do keep your workflow system under regular observation.

It is a good idea to review your working methods from time to time and ask yourself if it could be more effective. Even small misunderstandings and failures to learn from previous mistakes can have a major impact on hundreds of objects (e.g. incorrect resolution, wrong data formats, incorrect storage method, inefficient use of hardware and software).

Do avoid as much manual processing as possible.

It is worth checking carefully which digitization workflow tasks really need to be performed manually. Many recurring jobs can now be automated or at least partly automated using appropriate software. There are numerous tools from different providers covering image conversion, validation, image processing, metadata enrichment, OCR and many other tasks. Integrating these tools into your workflow as automated tasks not only saves time but also ensures that your results are consistent in terms of quality, data formats and naming systems.

Do prepare yourself for the technical challenges involved in all digitization projects.

Digitization projects involve numerous technical components. For this reason, at least one technician or a technically competent member of staff should be available to actively support the project, not only resolving problems as they arise but also maintaining computers and servers and acting as a point of contact for communication with hardware manufacturers, software producers, scanning service providers, and computer centers.

Do think about effective (long-term) archiving.

Reliable methods of long-term archiving can be expensive. In the ideal scenario, it may be possible to cooperate with other institutions and share the cost. If this is not an option, you will need to find effective alternatives. Archiving systems should be easy to maintain, and it must be possible to retrieve the archived data at any time. When choosing the most appropriate system, you should consider whether to store the data in two geographically separate locations.

Do get started!

While of course it makes sense to plan your project thoroughly and cover all eventualities, there comes a point when it makes equal sense to just get started. When taking that decision, you should consider the following points.

- Extracting and recording metadata is worthwhile, but only if the metadata is relevant and of some real benefit to the end user.
- Your choice of image quality (i.e. resolution) should take account of the visual capacity of the human eye. This principle also applies to compressed images and potential artefacts (keywords: visual lossless).
- Give some thought to the future maintenance requirements for the technical infrastructure and data formats so that any wrong decisions can be corrected at a later stage.

Do keep a hands-on approach to managing your project.

Project managers should work through each step of the workflow from start to finish at least once a year to ensure that they fully understand it, and if necessary ask the corresponding member of the project team to explain what is involved. This helps to disseminate and document knowledge across the entire project. It is also a good way to identify potential improvements and address typical pitfalls (e.g. "We've always done it that way").

Do establish links with a community.

Many other institutions will already have come across the same or similar problems. Establishing links with such institutions can help you to make the right decisions on procurement and working methods, exchange experiences, and learn from each other.

Don't create complex workflows with many side branches and loops.

Workflows should always be designed to run sequentially. Wherever possible, you should avoid repetition, branching and other complex procedures entirely. Especially if you are dealing with a large volume of data, simple and sequential workflows ensure that each step is easily traceable, making the task of identifying errors and maintaining your systems and workflows much easier.

Don't stick notes or labels onto old source material.

Barcodes, dockets, notes and other information should only be stuck onto old source material in exceptional cases. The chemicals used in the glue can seriously damage valuable objects. Ideally, any additional information should be provided on inserts or cover sheets made of acid-free paper.

Don't feel you have to do all the work in-house.

For every new digitization project, you should consider whether you really need to carry out the project in-house. Depending on the type of materials involved and the difficulties of working with them, it may make more sense to outsource the digitization process to an external scanning service and have the results delivered to you. This also avoids the need to buy expensive new hardware and additional personnel costs.
As well as the scanning process you can of course outsource other parts of the digitization workflow such as quality control, metadata enrichment, and OCR.

Scanning and photographing

Do validate your images automatically, ideally straight after scanning

Wherever technically possible, image files should be validated automatically as soon as they have been scanned, i.e. before performing any further tasks in the workflow. The free software JHove has proven effective for common methods of validation (e.g. of TIF files). It can be run automatically from a command line instruction and can return a machine-readable validation result in the form of an XML file.
It may also be worth validating file names at an early stage in the workflow. Software tools like 'regular expressions' can force project staff to comply with established naming schemes. They can also ensure that numerical sequences in file names are unbroken or that the file name includes relevant metadata.

Do think twice before deciding to use a scanning robot for your digitization project.

A scanning robot may not be the right choice for some types of source material. Tasks such as preparing the books, loading and unloading, and making fine adjustments for individual book formats can be relatively time-consuming depending on the model. Furthermore, not all books are equally suitable for scanning by a book robot. Experience has shown that current scanning robots are not yet reliable or autonomous enough to be able to operate entirely without supervision. The time required by a human supervisor to monitor the work of the robot and to load the device and make fine adjustments for each book can be as much as for manual scanning.

Do scan the color chart on each page as well

Ideally, each scan should include a color chart. This does not usually pose a problem for subsequent processing or when displaying the image as it can be detected and hidden through automated cropping.

Incidentally, color charts don't last forever and should occasionally be replaced.

Don't allow uneven lighting conditions in the scanning room to spoil your images

The brightness of the images produced by most scanners and cameras will vary depending on the ambient lighting conditions (summer/winter and morning/afternoon). Although they have their own integrated lighting systems, they are nevertheless influenced by other light sources in the room due to the way they are designed. Ideally, this should be thoroughly checked at least once for each device. Even hardware that is advertised as operating independently of ambient light can deliver surprising results under close examination of two photographs taken under different lighting conditions.
All images of the same object should be taken under constant lighting conditions.

Don't allow the scanning software to perform automatic cropping

The option to let your scanning software crop images automatically may sound attractive as a way of reducing your workload, but the potential benefits and disadvantages should be carefully weighed up for each project. If the images are cropped too heavily, the project may not be able – at some point in the future – to generate more information from them than was anticipated at the time of scanning. For this reason, especially in the case of valuable materials that should ideally not be exposed to multiple scanning, it is important to ensure that the master image is produced with a generous all-round margin.
The scanning program should never be instructed to automatically perform 'heavy cropping' of master images. Instead, an automated cropping step can be incorporated into the remaining workflow using a derivative image and selected parameters. This step can be

performed as many times as required without changing the master image.

Files and formats

Do check which is the central data system.

Before you start work on a digitization project, you should clarify which system is to be used for data storage. Is there a library catalogue, an archive database or something comparable? Should the data be enriched in this system before digitization, or should any metadata be extracted and recorded solely within the workflow?

Do use appropriate data formats.

All data formats used in digitization projects should be chosen carefully. Wherever possible, digitized output and metadata should be stored using standardized file formats that are well-established and suitable for machine processing. In exceptional cases, if you need to use your own proprietary data format, you should ensure that your files are nevertheless machine-readable, e.g. by using an XML file structure.

Do conduct pre-checks on data received from partners.

If you plan to work with a partner, you should ideally request and conduct pre-checks on a relatively large body of data before commencing the project. Actual data provide the only reliable indication of whether all the project partners have the same understanding of the delivery format. The later you receive and pre-check the data, the greater the risk that a large volume of data could be produced in the wrong format and therefore be unusable.

Do make judicious use of the available options for extracting and recording metadata.

Your approach to extracting and recording metadata should be judicious and pragmatic. Researchers using your digital collection will appreciate being able to locate and open a particular chapter on the basis of descriptive metadata. By contrast, few users will be interested in the distinction between a title, formal title, uniform title, heading, other titles, title proper and an equivalence. Equally, most users will be happy with the author's given name and family name, so there may well be no need to include additional name information such as titles, prefixes, numbers, nicknames, or generic names.

Do validate your metadata.

As a rule, all metadata should undergo an automated validation process. This should be performed as early as possible in the workflow to avoid potential problems at a later stage when working with other technical systems. Validation processes are very simple to configure on the basis of "regular expressions".

Do wait until your images have been cropped before performing OCR.

As a general rule, OCR should not be performed until your images have been cropped, black borders removed, the book-fold area trimmed, and the images deskewed or scaled. Otherwise there is a risk that the full text and the corresponding word coordinates (e.g. in the output ALTO files) will no longer match the modified image coordinates.

Do store raw data whenever possible.

The raw data produced in the course of individual workflow steps should be stored whenever it is feasible and useful to do so. By way of example, the original OCR raw data format can be used at a later stage to provide further information to supplement that available from the converted data, e.g. information about the layout, fonts, languages, recognition accuracy rates, and statistics. Even the master digitized images can be used later on to extract further information.

Do publish your results.

In the ideal scenario, digitization results should be made publicly available. The OAI-PHM and SRU interfaces provide established standards that allow other users, harvesters and portals to make use of your data. External partners can also work with the data that you supply or enrich it using crowdsourcing platforms.

Don't make up new metadata standards.

Only existing and open standard formats should be used for descriptive metadata. Avoid creating your own new document types and metadata definitions. The MARC and MODS standards of the Library of Congress are generally adequate for most purposes.
Under no circumstances should the end product of your digitization workflow be in a proprietary format. Wherever possible, you should work with established formats such as METS/MODS, LIDO and EAD for descriptive metadata and ALTO or TEI for full text.

Don't expect your end users to be metadata experts.

Don't assume that the end users of you digitized material will have anything more than a superficial knowledge of metadata. The contents of your online digital collections do not need to meet the stringent requirements of a library catalogue. For most users, it is more than enough to provide a link from the digitized object to the full catalog entry. This streamlined approach will save a great deal of workflow time otherwise spent manually extracting and recording metadata.

Hardware and software

Do use open-source tools.

As a general rule, open-source tools are preferable as they encourage a longer-term approach to the use of digitized content and software components. This also allows you to modify or supplement the software independently of the manufacturer without having to switch to another solution.

Do include scalable hardware in your planning.

Key elements of the overall technical system should be scalable. Virtualization, for example, makes it easier to scale up your hardware resources. It is also worth considering right from the start of your project whether to outsource some of the more CPU-intensive workflow steps such as data conversion. When choosing the file system, you should ensure that it is efficiently structured so that it can deal with large volumes of data and so that the amount of storage can be increased flexibly.
System resources should be monitored actively to prevent disruption and identify scaling requirements in good time.

Do conduct regular checks on back-up and archiving procedures and on the actual back-up files.

Back-up and archiving procedures should be checked at regular intervals. Ideally, these checks should be performed in line with a regular schedule (e.g. every six months).
At the same time as you check that a back-up copy has been correctly produced, you should also check that it can be correctly retrieved, read and processed.

Don't be tempted to rush out and buy new scanning hardware.

Not everything the hardware manufacturers tell you during their sales routine will be suitable for your project or source material. You should always think very carefully before investing in new scanning equipment and ideally test it using your own source material before purchase. In particular, you should closely examine features such as the device's light dependency, resolution, opening angle, ergonomic design, and efficiency. If possible, arrange to put the hardware through its paces in your own scanning room.

Don't regard your own computer as a safe location to store your data.

Your desktop PC is not a secure place to store data. As well as potential hardware problems, your data is at risk from viruses and user errors. For these reasons, you should always store your expensively produced master images and all your back-ups on a professionally maintained server with a back-up system.

Don't use central network drives for which a large number of people have write permission.

Avoid using network drives for which a large number of users have simultaneous write permission. Otherwise, if your data is lost, it may not be possible later on to identify how the loss occurred and how similar losses can be prevented.

Don't use external hard drives as secure storage.

External hard drives and USB memory sticks are not suitable for backing-up or archiving purposes. Every digitization project should have a robust and properly designed infrastructure to ensure that the output data are permanently and securely stored.
Bear in mind that it may not be possible to redigitize the same object. You should therefore ensure that your digital copy of the object is safely and permanently stored. It may even be worth storing the data twice in separate locations.

Examples such as these, as well as tips on best practice, commonly used software tools and even the choice of project furniture can have a significant impact on productivity in digitisation projects. Above all, however, it is a combination of workflow structure, usability and smartly implemented (semi-) automation that helps to boost throughput (in some cases by as much as 100%) without deploying additional hardware or staff.

Author Biography

Steffen Hankiewicz is a senior software developer, CEO and owner of the German software company intranda GmbH. He has been developing and implementing software solutions for digitization projects for more than 16 years. The open-source workflow management software Goobi, the Goobi viewer as well as an automated TaskManager for OCR, JPEG2000 conversion or Named Entity Recognition jobs are some of the current digitization tools he develops and supports together with his team.