

# Best Practices Digitization

(Version 1.1, 2016)

## Content

|  |    |
|--|----|
| 1. Planning.....                                 | 2  |
| 1.1. Project Scope.....                          | 2  |
| 1.2. Collaborations.....                         | 3  |
| 1.3. Cost Factors and Funding Forms.....         | 3  |
| 1.4. Own Production or Outsourcing.....          | 5  |
| 1.5. Legal Aspects.....                          | 6  |
| 2. Preparation of the Material.....              | 7  |
| 3. Digitization Process.....                     | 8  |
| 3.1. Scanners.....                               | 8  |
| 3.2. Digitization Parameters.....                | 12 |
| 3.3. Quality Control and Image Processing.....   | 14 |
| 3.4. Full-Text Digitization.....                 | 16 |
| 3.5. File Names.....                             | 16 |
| 3.6. Digitization Workflow.....                  | 18 |
| 4. Storage, Archiving.....                       | 19 |
| 4.1. File Formats, Master and Usage Formats..... | 19 |
| 4.2. Memory and Long-Term Archiving.....         | 20 |
| 5. Metadata, Recording Instruments.....          | 21 |
| 5.1. Interfaces for Metadata.....                | 22 |
| 5.2. Recording Digitization Projects.....        | 22 |
| 6. Sources.....                                  | 23 |
| 6.1. General Works.....                          | 23 |
| 6.2. Digitization Projects and Platforms.....    | 24 |
| 6.3. Service-Providers.....                      | 25 |

# 1. Planning

Project risks, such as funding problems, poor quality of the results, legal barriers or inefficient processes, can be minimised with some careful planning and conceptualisation. This website explains some key points for the conception of a digitization project.

## 1.1. Project Scope

### Partial or entire holdings?

Modern scanning technology, falling storage costs and software that simplifies the workflow have triggered a paradigm shift in digitization concepts: instead of individual documents, entire collections are being digitised. Old holdings in the so-called public domain are prioritised here as the copyrights have run out for them. Prime examples of mass digitization are [Google's](#) digitization project or projects like [e-rara.ch](#) or [E-Pics](#), which ETH-Bibliothek is in charge of coordinating.

Selection criteria for the digitization of holdings might be:

- Research relevance
- (Academic) demand
- Preservation of holdings (frequently used materials, unique specimens)
- Virtual reconstruction of collections and holdings

#### Practical example – project scope E-Periodica

The materials for the digitization project [E-Periodica](#) are selected according to the following criteria:

- Journal character: typical journals that appear more frequently than once a year take priority; in addition to this, yearbooks and other series
- Reference to Switzerland based on content, authorship and publisher, language, place of publication etc.
- Non-commercial character: non-profit publishers (e.g. specialist associations) or small, local publishing houses
- Support from publishers: for journals that are still published, digitization is considered only with the active support of the publisher
- Addition to an existing digital journal collection

### Volume framework

A detailed volume framework offers a decent basis for the project planning. Especially for heterogeneous holdings, it can be helpful to record the number, type and nature of the materials.

For written documents, such as books and journals, it is an advantage if the number of pages is known or can be estimated as accurately as possible.

## Memory space

For extensive projects, data quantities can soon run into the terabyte range. Therefore, it is important to plan ahead where and in which form the data is to be stored or archived (see also [Follow-Up Costs](#)).

## 1.2. Collaborations

Within the scope of collaborations, content-related aspects can be coordinated, redundancies avoided and costs shared. A reduction in costs can be achieved through the joint procurement and/or usage of the scanner infrastructure.

Important aspects of a cooperative agreement:

- Project scope and holdings
- Project duration and organisation
- Services of the project partners
- Granting of rights
- Funding
- Guarantees and liability
- Duration and termination of the agreement
- Costs and cost-sharing

The early exchange of information on digitization proposals and the use of central reference instruments is generally important.

### Practical example – collaborations e-rara.ch and e-manuscripta.ch

Both [e-rara.ch](#) and [e-manuscripta.ch](#) are conceived and structured as joint projects between Swiss university libraries. Meanwhile, however, numerous other libraries and institutions are also involved and contribute towards the continuous expansion of the service with their holdings.

## 1.3. Cost Factors and Funding Forms

The main cost drivers for digitization projects are personnel expenses and the development of the technical infrastructure and knowhow.

## Cost factors in general

- Personnel expenses for scanning, quality control, preparation and reworking of the material, possibly training costs on the scanner
- Infrastructural costs (e.g. scanner, software solutions, databases, IT support)
- Licensing costs for image processing and indexing programmes, full-text recognition (OCR)
- Nature of the original (bound, loose-leaf, special formats, handling)
- Scanning parameters (colour/greyscale, resolution/dpi)
- Online presentation (IT infrastructure platform, support, metadata recording)
- Transport and insurance premiums

## Cost factors for external digitization service-providers

Outsourcing to external service-providers especially pays off for an individual project with a volume that would not justify the internal procurement costs. Depending on the nature and scope of the projects, it is advisable to obtain two to three quotes.

## Follow-up costs

The follow-up costs for the maintenance of software, hardware and data depend on the individual case and usually can only be estimated roughly.

The same goes for storage and long-term archiving expenses. Although the prices for storage capacities have been falling steadily in recent years, they can still stretch the project budget for large digitization endeavours. The fact that it involves long-term, running costs beyond the project framework needs to be considered here. Moreover, the greater the data quantity, the higher the cost of data protection and subsequent data migrations.

## Funding forms

There are different funding possibilities, depending on the project and its framework conditions:

- Self-funding
- [Collaborations](#)
- Third-party funding (e.g. support programmes)
- Public Private Partnership: cultural institutions collaborate with private companies and jointly perform public services, e.g. the cooperation between the [Bavarian State Library](#) and [Google](#).
- Sponsoring

### Practical example – funding E-Periodica

Specialist associations and publishers usually only have limited means at their disposal to fund digitization projects. In the early days, E-Periodica (previously retro.seals.ch) was co-funded with project money from [e-lib.ch](http://e-lib.ch) and using the [Consortium of Swiss Academic Libraries](#) and ETH-Bibliothek's own funding. Currently, the cost model looks as follows:

- One-off costs for the processing and activation of the retro holdings for a journal: funded by the project partners (publisher, specialist association etc.) and ETH-Bibliothek
- Annual costs for the addition of new issues and memory, and operating costs for the entire journal are passed on to the project partners in full  
*The umbrella organisation [Swiss Academy of Humanities and Social Sciences \(SAHS\)](#) contributes towards funding the annual costs in journals for the humanities and social sciences.*

## 1.4. Own Production or Outsourcing

Digitization projects require knowhow and modern technology. The individual requirements, project scope and resources available (personnel, funding, infrastructure etc.) determine whether and to what degree external service-providers should be used.

The following table provides an overview of the main pros and cons of in-house digitization and outsourcing:

| Pros outsourcing  | Cons outsourcing  |
|---|---|
| <ul style="list-style-type: none"> <li>• Low (own) staffing requirement</li> <li>• No specialist, in-depth knowledge needed for the scanning process itself</li> <li>• No personnel resources needed for the scanning process</li> <li>• Low level of investment necessary</li> </ul> | <ul style="list-style-type: none"> <li>• Intensive work preparation</li> <li>• Complicated prior clarifications (quotes, test runs)</li> <li>• Dependence on external delivery times (risk of project delays)</li> <li>• Logistics overheads (transport, insurance, checks etc.)</li> </ul> |
| Pros in-house   | Cons in-house   |
| <ul style="list-style-type: none"> <li>• Independence (deadlines etc.)</li> <li>• Simple planning</li> <li>• Lower logistics overheads</li> <li>• Short communication channels</li> </ul>   | <ul style="list-style-type: none"> <li>• Commitment of personnel resources</li> <li>• Specialist knowledge needed</li> </ul>  |

- |   |  |
|---|--|
| <ul style="list-style-type: none"><li>• Experience and learning options</li></ul> | <ul style="list-style-type: none"><li>• New challenges (resource planning, new infrastructures, storage management etc.)</li></ul> |
|---|--|

Outsourcing does not necessarily mean “off-site”: for larger projects or holdings that cannot be taken off-site, service-providers who work on-site with their own equipment and staff can be used.

## 1.5. Legal Aspects

The legal situation should be verified and usage rights obtained as early as possible during the planning phase.

### Copyrights

Under Swiss [copyright law, \(pdf, 255 kB\)](#), works are protected until seventy years after the author’s death, after which they are freely available as so-called “public domains”. A work published in 1900, the author of which died in 1950, is therefore under copyright until 2020.

Whether the publisher of a work holds the rights for its (additional) publication on the internet needs to be verified. For older publications, this can be excluded from the outset as the internet did not exist when the contract between the author and the publisher was agreed.

### Artistic works: ProLitteris

If the originals are artistic works or images, permission for their online publication needs to be obtained from [ProLitteris](#), which represents the rights of the participating authors and photographers and charges a (fixed) usage fee on their behalf for an additional online publication. These additional costs also have to be taken into account and who will pay for them in the long term needs to be clarified.

#### Practical example – copyright e-rara.ch

Apart from a few exceptions, primarily old prints that were published before 1900 and where no copyright restrictions apply anymore are uploaded onto e-rara.ch.

### Practical example – copyright E-Periodica

The prerequisite for an upload onto E-Periodica is the involvement and consent of a publication's copyright holders. An exception is made for journals until the 19<sup>th</sup> century, where the copyright holders can no longer be ascertained.

Experience has shown that only the publishers of newer commercial journals conclude contracts with their authors. In doing so, the authors transfer the full usage rights to the article in question to the publisher (copyright agreement). For historical contents, such clear agreements usually do not exist and obtaining retroactive permission from authors and their descendants through the correct legal channels would involve a disproportionate amount of effort.

#### *Dispensation with formal permission*

For journal digitization at E-Periodica, a pragmatic procedure has therefore been adopted: the author's consent is not obtained, but the author can subsequently demand the blockage or deletion of his or her work from the online service.

This approach assumes that the copyright holders are academically active people or part-time authors who support the further dissemination of their work in the interests of free science.

## 2. Preparation of the Material

In order to ensure that a project goes as smoothly as possible, issues concerning the preparation and handling of holdings should be clarified in as much detail as possible, especially in the case of mass digitization.

### Indexing

If a catalogue entry, database or suchlike is not already available, the holdings need to be indexed with essential [metadata](#). For smaller holdings, an Excel table may be sufficient.

Every document requires an identifier for the clear identification of a work, based upon which the scans can be assigned to the respective dataset. This might be the individual system number on the library catalogue, for instance (see also [File Names](#)).

### Preparation of the material

Depending on the type of holdings, the preparation of the material may be more or less complicated.

Possible preparatory work might be:

- Cutting-up of doublet holdings that are no longer needed (e.g. journals) by a bookbinder

- Document labelling
- Removal of staples
- Pre-sorting, formation of digitization units

### Compiling the documentation for the digitization process

Detailed information on the holdings facilitates the planning of personnel and financial resources. The necessary information can be presented in a volume framework.

- Different materials have different digitization requirements. The important parameters include the sort of paper, pages in special formats, fold-out sections, colour pages, figures, drawing etc.
- For valuable holdings, documentation on the state of the works may also be helpful to record any existing damage, for instance.

#### Practical example – preparation e-manuscripta.ch

- The documents (e.g. letters, dossiers, notebooks, sketches) are recorded in an archive database
- Staples, folders etc. are removed
- Documents that belong together (e.g. a letter with an envelope) are combined into units with the aid of an acid-free envelope
- Every unit receives a cover sheet with an identification number, which is generated from the existing database. This cover sheet contains the most important information: title, number of pages to be scanned

## 3. Digitization Process

A digitization process is centred on the use of suitable hard- and software for the scanning process itself, but also for subsequent quality control or image processing. However, the later traceability of the contents is also a key aspect, which can be facilitated by providing metadata and full-text files. Process-supporting tools based on workflow software are available for extensive digitization projects.

### 3.1. Scanners

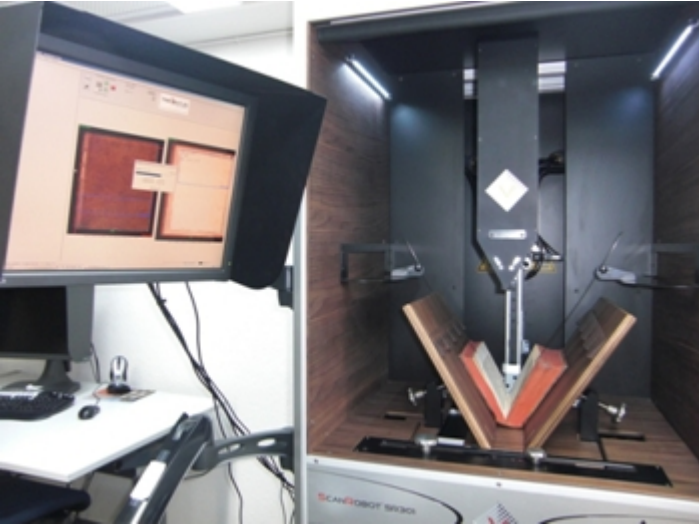
Which type of scanner is suitable for a particular project depends on the nature of the original document (type, condition, format etc.) and the scope of the project.



For sensitive documents, the choice of equipment and components should be coordinated with an expert.

The most important scanner types at a glance:

| Scanner   | Description, Pros  |
|---|--|
| <p><b>Document scanner</b></p>                                  | <ul style="list-style-type: none"> <li>• Scanning of individual, unbound pages</li> <li>• Extremely rapid processing with a high efficiency level (up to approx.. 50 pages/minute), automatic feeder</li> <li>• Software enables further processing steps (e.g. conversion into PDF)</li> <li>• Simple to operate</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Cutting-up of the documents</li> <li>• Limited scan quality</li> <li>• Not suitable for documents that are challenging from a conservational perspective</li> <li>• Errors possible in the form of a double page feed</li> <li>• Adverse effect on the reader unit through dust particles (can deteriorate over time)</li> </ul> |
| <p><b>Reflected light scanners<br/>(also: book scanner)</b></p> | <ul style="list-style-type: none"> <li>• Scanning of bound documents</li> <li>• Style sheets up to A0 (depending on type of scanner)</li> <li>• Book cradle to help position the book optimally</li> <li>• Special attachments enable gentle scanning at an opening angle of up to (min.) 90 degrees</li> <li>• Simple to operate</li> <li>• Simultaneous creation of different <a href="#">derivatives</a> (e.g. JPEG) besides the master file (TIFF)</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Limited throughput due to manual operation</li> <li>• When working without an attachment, the document needs to be open 180 degrees</li> </ul>  |

|  |  |
|--|--|
|  | <ul style="list-style-type: none"> <li>• The attachment slows down the scanning process</li> </ul>   |
| <b>Scanning robots</b>                                     | <ul style="list-style-type: none"> <li>• Very high throughput</li> <li>• Gentle processing (small opening angle)</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Only suitable for certain books. Limiting criteria include format or nature of the paper</li> <li>• Book covers and fold-outs cannot be scanned with a scanning robot; these need to be processed separately on a reflected light scanner</li> </ul>  <p>Treventus ScanRobot</p> |
| <b>Special scanners</b>                                    | <ul style="list-style-type: none"> <li>• Special constructions such as the <a href="#">Grazer Book Table</a> or the <a href="#">Wolfenbuettel Book Reflector</a> enable a gentle, non-contact digitization of fragile documents (e.g. manuscripts, early printed books, codices)</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Complicated individual settings and manual handling result in a low throughput</li> </ul>   |
| <b>High-quality digital cameras for image digitization</b> | <ul style="list-style-type: none"> <li>• For image documents: photographs, postcards, slides, negatives</li> <li>• Suitable for high quality demands</li> <li>• Suitable for mass digitization</li> </ul>  |

|                               |   |
|-------------------------------|---|
|                               | <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Technical or photographic knowledge necessary</li> </ul>   |
| <b>Film scanners</b>          | <ul style="list-style-type: none"> <li>• For negatives in a 35mm or medium format and slides</li> <li>• Suitable for mass digitization, batch-based processing</li> <li>• Simple to operate</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Sometimes limited scanning quality</li> </ul>   |
| <b>3D systems for objects</b> | <ul style="list-style-type: none"> <li>• <b>3D camera system (360° object representation):</b> consecutive images of an object in a 360° view (uni- or biaxial). And subsequent assemblage of the images by a 3D viewer</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Time-consuming, depending on the type of object</li> <li>• Quality of the result also depends on the viewer's technical prerequisites</li> </ul> <ul style="list-style-type: none"> <li>• <b>Laser technology:</b> creation of a so-called point cloud with central survey points. This point cloud is the starter model for the virtual reconstruction of the object</li> </ul> <p><i>Cons:</i></p> <ul style="list-style-type: none"> <li>• Very complex system, not suitable for mass production</li> <li>• Requires specific technical knowhow</li> </ul> |

### Practical example – 3D camera system e-pics.ch

Presentation [ETH E-Pics Earth Science Collections](#)

## Selecting the right device

A list of the main criteria for selecting a scanner is presented below. Tests with your own material and standardised test images are recommended to verify the quality. These are usually provided by the device manufacturers. Test phases in your own house can also be agreed with good providers.

Criteria for the assessment of scanners:

- Image quality: sharpness, colour, resolution, use of ICC profiles
- Protection of the documents: scanning without glass, harmless light, opening angle
- Scanning speed in practical use
- Handling, flexibility (different application possibilities for an optimum workload)
- Software, output format, possibility for network and storage connection
- Transportability, ergonomics
- Spatial conditions: areas, ceiling height
- Noise emissions
- References, support, maintenance
- Costs

### 3.2. Digitization Parameters

A master file is the starting point for all further processing of digital copies. It is used to generate JPEG files, for instance. The most important parameters for the production of the master file are explained in more detail in the following section:

#### Image resolution

Resolution refers to the number of pixels per unit of length. Its unit of mass is dpi (dots per inch) or ppi (pixels per inch). The higher the resolution, the more detailed the digital copy and the greater the volume of data.

*General recommendations:*

- 300 dpi for greyscale and colour originals
- 400 dpi for special manuscripts, prints or maps with delicate content
- 600 dpi for bitonal scans (black-and-white originals)
- Higher resolutions: only make sense for special applications, such as studying paper structures or image digitization (3,000 to 4,000 dpi for negatives and slides)

## Colour management

The colour reproduction of an image can differ greatly from one image and reproduction device to the next, such as scanner, image monitors and printers. The [Colour Management System \(pdf, 1.28 MB\)](#) guarantees an identical colour reproduction, regardless of the imager and output device used. Calibrating the devices with a standardised colour profile can generally even out any colour differences. As a rule, the ISO-certified [ICC Profile](#) by the International Color Consortium is used. The colour profile deployed is stored together with the digital copy.

For originals where colour is a key criterion for research questions, the inclusion of an additional standardised colour scale (or colour checker) along with the original is recommended.

### Practical example – colour management e-rara.ch

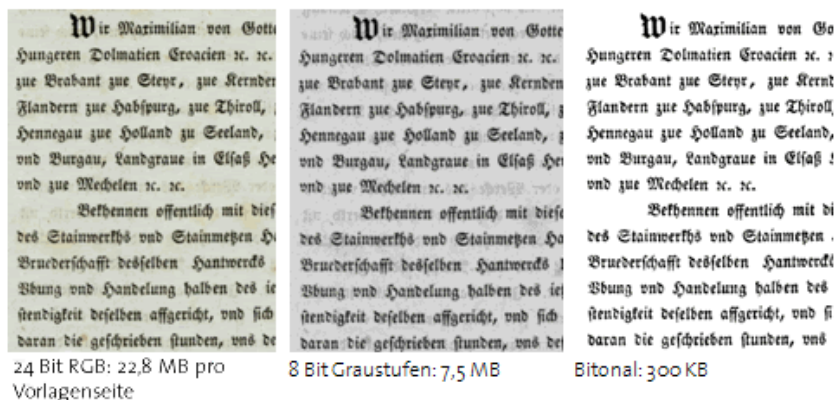
In the project e-rara.ch, every fourth page is scanned together with a colour scale:



However, this scan is technically disabled and not visible in the digital work online.

## Colour intensity

Colour intensity refers to the amount of the colours that can be represented. It is indicated as the number of available bits per pixel. A high colour intensity facilitates fine colour shades, but requires more memory and incurs higher storage costs.



Major differences: a scan in colour requires three times as much memory as the original in 8-bit greyscales and many times the bitonal version.

The intended use of the digital copy is important for the colour intensity required:

- For manuscripts and old prints (until approx. 1750), colour digitization is recommended (German Research Foundation)
- For later prints or books with (non-colour) illustrations, a greyscale scan is usually sufficient
- For the mass digitization of simple, illustration-free prints from the 19<sup>th</sup> and 20<sup>th</sup> centuries, black-and-white – so-called bitonal – digitization is usually suitable

### 3.3. Quality Control and Image Processing

Consistent quality control anchored in the workflow is vital, i.e.:

- Minimum requirements should be defined for quality control on a project basis
- If possible, quality control is not performed by the scan operators themselves, but rather a second person as sources of error are more likely to be discovered with two pairs of eyes

#### Control through workflow software

Although current software solutions to support workflow usually perform a quality control autonomously, this is limited to technical aspects (file format, resolution, colour management etc.). It is not a complete substitute for a visual control.

Typical scanning errors:

- Missing or double pages
- Shadowing or finger marks on the digital image
- Slanting pages
- Cropped type area (text block on the page)
- Insufficient image definition
- Poor colour authenticity
- Image interference (e.g. moiré effect)

#### Practical example – quality control E-Periodica

The journals scanned for E-Periodica are usually checked twice:

1. The completeness, legibility and quality of the image files are checked immediately after scanning. Standard tools such as [Adobe Bridge](#) are used on calibrated monitors.

2. An additional visual control is performed while the digital copies are being indexed with metadata.

From experience, errors that are only discovered during metadata entry take a great deal of effort to correct: rescanning, replacement of the master file, replacements of the [usage derivatives](#), recovery of the correct file name etc. Therefore, the initial manual control is of paramount importance.

## Image processing

Retrospective image processing is time-consuming and costly. Especially in the course of mass digitization, there is no guarantee that the effort involved in individual image processing will pay off. In order to ensure an adequate image quality, however, there is software that enables batch image processing based on pre-set standards. Examples are [PageImprover](#) by 4Digital Books or the open-source product [ImageMagick](#).

### Practical example – image processing E-Periodica

#### Software

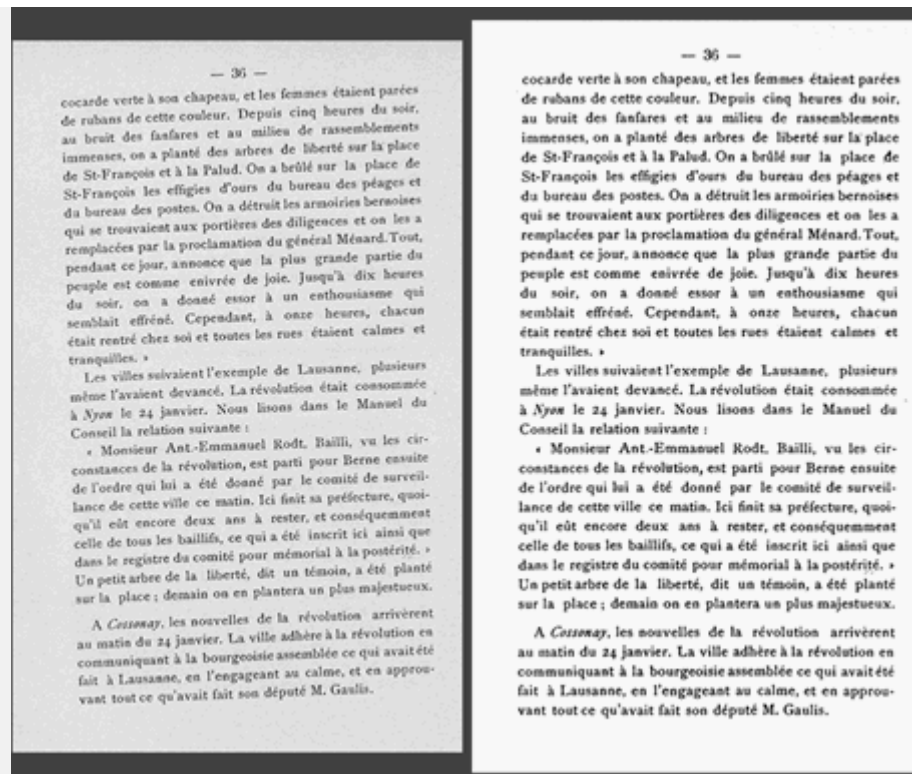
The scanners used by ETH-Bibliothek have device-specific, multi-purpose image processing software. This does not only have advantages for the digitization of journals: as the processing possibilities differ from one scanner to the next and not every device can satisfy every need, a separate professional image processing software programme (PageImprover) is used for subsequent processing.

#### *Semi-automatic image processing*

The image processing usually takes place in batch mode. The optimisation parameters are determined manually in advance based on individual pages; the automatic image processing is ultimately performed based on these settings.

The concrete optimisation tasks in journal digitization:

- Alignment of the document
- Reduction of disruptive background information (translucent pages)
- Increase in the contrast between the text and the background in greyscale scans
- Page-centring



The original scan (left) compared to the optimised image file (right).

### 3.4. Full-Text Digitization

In order to create a full-text document, either the original needs to be copied down manually or an OCR software programme can be used. OCR stands for “Optical Character Recognition”. It recognises letters or characters in image files and enables them to be used as text.

OCR is primarily suitable for more recent Antiqua script. For the majority of historic materials, however, OCR can only be applied to good effect in individual cases— if at all. It is generally unsuitable for the recognition of manuscripts. German type can sometimes be recognised with special software solutions, e.g. [ABBYY Finereader XIX](#).

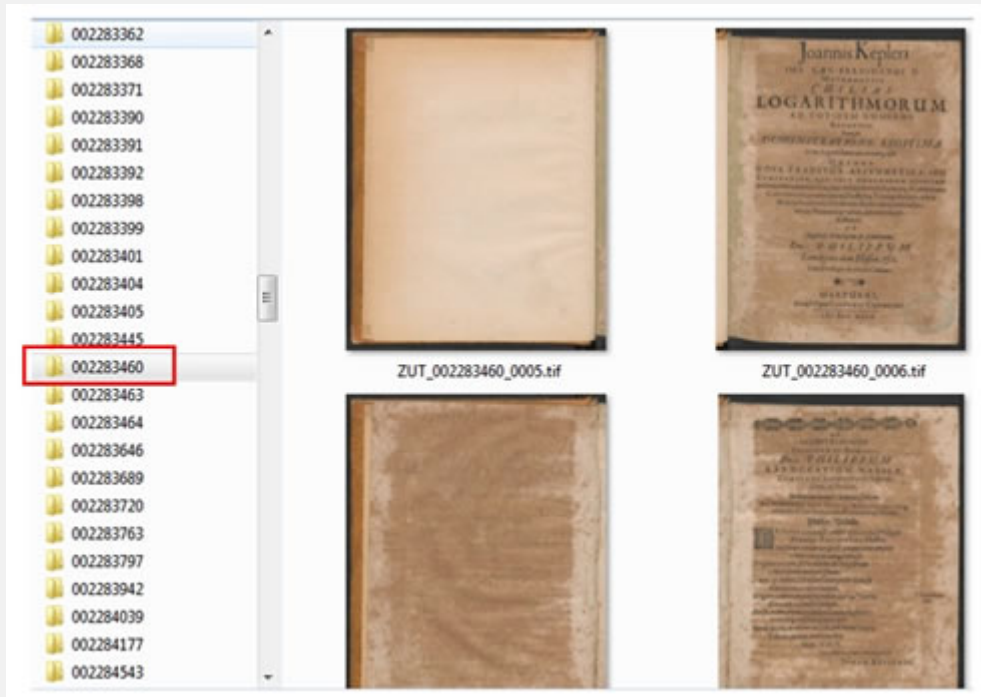
### 3.5. File Names

The digitised documents need to have clear file names and be part of a persistent and clear thread. Numbering files consecutively with a serial number is recommended. If possible, the file names should refer to the content.



## Practical example – file name e-rara.ch

### 1. Name of the file with the individual scans



### 2. Clear identification of the related catalogue dataset

|                          |   |
|--------------------------|---|
| <b>Titel</b>             | <a href="#">Joannis Kepleri ... Chilias logarithmorum ad totidem numeros rotundos quibus nova ! admirabilis, nec utilis solvendi ... : praemissa demonstratione legitima ortus logarit[ue] quo post numerorum notitiam nullum nec admirabilis, nec utilis solvendi pleraq[ue] divisionis, radicumq[ue] extractionis, in numeris prolixis, labores molestissimos ...</a> |
| <b>Impressum</b>         | <a href="#">Marpurgi : excusa typis Casparis Chemlini, 1624</a>   |
| <b>Umfang</b>            | 55 S., [26] Bl., [1] gef. Bl., S. 113-216 : Ill. ; 19 cm (4°)   |
| <b>Enthalten in</b>      | <a href="#">Sammelband ETH-BIB Rara</a>   |
| <b>Gehe zu</b>           | <a href="#">Sammelband ETH-BIB Rara</a>   |
| <b>Bibl. Nachweis</b>    | VD17 23-254285W   |
| <b>Titelvariante</b>     | <a href="#">Chilias logarithmorum ad totidem numeros rotundos quibus nova traditur arithmetica, utilius solvendi ...</a>  |
| <b>Inhalt:</b>           | Supplementum chiliadis logarithmorum, continens praecepta de eorum usu ... 1625   |
| <b>Gesamtbestand</b>     | <a href="#">Alle Exemplare</a>  |
| <b>Autor/-in</b>         | <a href="#">Kepler, Johannes, Astronom, Mathematiker, 1571-1630. ger</a>  |
| <b>Druckort</b>          | <a href="#">Marburg (Lahn)</a>  |
| <b>Druckerei/Drucker</b> | <a href="#">Chemlin (Offizin, Marburg)</a>  |
| <b>Systemnr.</b>         | <b>002283460</b>  |

### 3. Automatic import and display of the metadata in the presentation interface



| Titelaufnahme   |  |
|---|--|
| Titel   | Joannis Kepleri ... Chiliae logarithmorum ad totidem numerorum notitiam nullum nec admirabilis, nec ut [ue] usus quibus nova traditur arithmetica, seu com solvendi pleraq[ue] problemata calculatoria, praeser extractionis, in numeris prolixis, labores molestissim |
| Autor, Beteiligte   | <a href="#">Kepler, Johannes</a>   |
| Impressum   | <a href="#">Marpurgi : excusa typis Casparis Chemlini, 1624</a>  |
| Umfang  | 55 S., [26] Bl., [1] gef. Bl., S. 113-216 Ill. 19 cm (4 <sup>o</sup> )   |
| Bibl. Referenz  | VD17 23:254285W  |
| Sprache   | Latein   |
| Standort des Druckexemplars   | ETH-Bibliothek Zürich, <a href="#">Rar 5357</a>   |
| Persistent Identifier (DOI)   | <a href="https://doi.org/10.3931/e-rara-4650">10.3931/e-rara-4650</a>   |
| Kollektion  |  |
| <a href="#">Weitere Kollektionen</a> » <a href="#">Mathematik und Physik (ETH-Bibliothek)</a> |  |
| Inhalt  |  |
| <a href="#">Inhalt des Werkes</a>   |  |

### 3.6. Digitization Workflow

Workflow systems aid the efficient and effective organisation of the complex digitization process. The allocation and systematics of file names is one aspect of the workflow. This section outlines possibilities of automating individual work steps in the digitization process.

#### Workflow software

Software solutions support the workflow from scanning and importing data to providing it on the internet. Even mass digitization projects can be conducted and managed efficiently with their aid.

Workflow systems are structured modularly and governed by the typical project procedure. The individual tools and functions can be subdivided broadly into indexing and management, and presentation.

For instance, the software [Goobi](#) is commonplace in the library sector.

#### Practical example – workflow software E-Periodica

The journal projects by E-Periodica are carried out with the software Agora by the company [SRZ Berlin](#), which offers various modules for the management of heterogeneous data types:

- Agora Process (converter): Windows service for batch-based image processing and format conversion, as well as interface with OCR engines
- Agora XML Editor: recording bibliographical and structural metadata

- Agora Production Repository: storage and indexing of object information, such as full text, structural and metadata
- Online Repository: tool for presentation on the internet

#### Practical example – workflow software e-rara.ch and e-manuscripta.ch

The two aforementioned projects are being conducted based on the [Software Visual Library](#), produced by [semantics](#) and provided by the company [Walter Nagel](#).

Based on this multi-client-enabled platform, the following steps can be processed:

- Import of the digital copies from the libraries
- Automatic ingestion of the corresponding metadata from the various catalogue systems
- Automatic conversion into the display format JPEG and processing for the zoom view
- Partially automated quality assurance (not currently used in either project).
- Recording of the structural and pagination data
- Online presentation on the native web portals
- Creation of search filters based on facets, lists and clouds
- Automatic dispatch of archive capsules to local sites

## 4. Storage, Archiving

### 4.1. File Formats, Master and Usage Formats

Digital images can be stored in various file formats. Every format has properties that affect the usage possibilities (presentation on the internet, long-term archiving etc.). Moreover, the memory required depends on the file format selected.

Some file formats reduce the file sizes by compressing the images. In doing so, a distinction is drawn between lossless and lossy compression. Normally, lossy compression cannot be undone – in other words, the original file cannot be reconstructed after compression.

Common file formats used in digitization: TIFF, JPEG, JPEG2000, PDF, PDF/A.

A detailed description of file formats is provided by the *Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen* ([KOST](#)).

## Master and usage formats

### Master file

A master file can be described as the unprocessed scanning product or original file. Master files of greyscale and colour digital copies are usually stored as uncompressed TIFF files.

As an alternative to TIFF, JPEG 2000 requires less memory and, unlike the JPEG format, supports lossless colour space compression and a progressive image build-up. Moreover, it can record metadata. In the [DFG Practical Guidelines on Digitisation \(2013, pdf, 761 kB\)](#), however, JPEG 2000 is not recommended as a storage format for master files – mainly due to the fact that (as yet) the file format is not very widespread.

### Usage derivatives

Derivatives are file products which have been optimised for use based on master files that are especially needed to display digital copies on the internet. To this end, compressed files are produced from the master files, which take up less memory and can be loaded quickly by the browsers. The browser-friendly formats JPEG and PNG are especially suitable for internet applications. For download functions or additional archiving purposes, the use of PDF/A is recommended.

#### Practical example – usage derivatives E-Periodica

For presentation on the internet, two derivatives are required in a JPEG format: one version in a reduced width for the overall view and one version in its original size for the zoom function. These are created automatically with Agora Workflow Client.

## 4.2. Memory and Long-Term Archiving

On average, a colour TIFF file with a resolution of 300 dpi requires 25 MB of memory space per scanned page. This means that only around 180 books can be stored on a storage medium with 1 TB of memory, for instance.

Essentially, two processes can be adopted to reduce data volumes:

- Storing the files (master and usage derivatives) in a compressed form
- Optimising the file size through suitable scanning parameters (resolution, colour intensity etc.)

## Long-term digital archiving

The conceptual planning for the long-term storage of original and master files is a central aspect of digitization projects. Cooperative solutions (e.g. with an efficient IT infrastructure) may help to conceive long-term archiving efficiently and safely.

Further information on the topic: [Digital Data Curation at ETH Zurich](#)

## 5. Metadata, Recording Instruments

There are three basic kinds of metadata:

- Bibliographic metadata, which describes the document and therefore is also referred to as descriptive metadata
- Structural metadata, which displays the document structure like a table of contents
- Administrative metadata, which provides information on usage rights, for example
- Technical metadata, which contains the technical parameters of a digital copy and provides information on the file type, file size, resolution etc.

For the practical realisation of a digitization project, descriptive and structural metadata that enables the document to be used practically in the first place are especially of paramount importance.

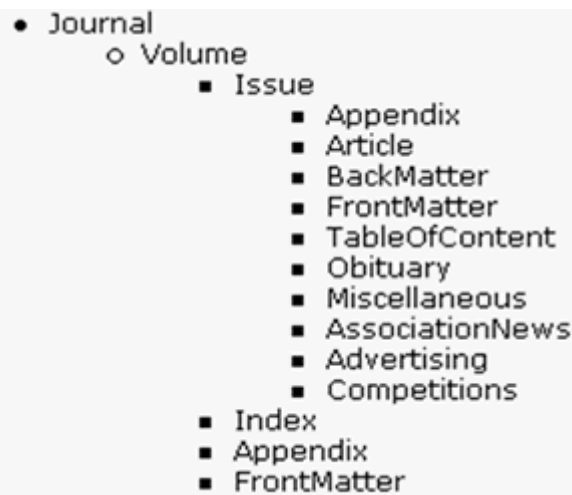
### Practical example – structural metadata E-Periodica

Bibliographical and structural metadata is recorded for journal projects by E-Periodica.

#### Structural elements

The document structure contains elements (e.g. journal, volume) and corresponding sub-elements (e.g. journal volume, issue, index, appendix and volume front matter), as well as attributes for these elements (author and title of the article). The precise application of the elements, sub-elements and corresponding attributes is stipulated in indexing rules created specifically for each journal based on the rules of formal cataloguing.

Content-related structuring takes place based on an XML editor, which is contained in the modular structure of the Agora workflow software. In structuring the content, the page number of the printed number is assigned to each digital copy.



Example of a document structure

## 5.1. Interfaces for Metadata

Metadata can be exchanged between systems, provided that the data complies with international standards.

For this purpose, the platform should provide its data in [Dublin Core format](#) and use OAI-PMH ([OAI-Protocol for Metadata Harvesting](#)). OAI-PMH also supports XML-based metadata formats. According to the German Research Foundation, metadata should also be available in [METS](#).

### Interfaces for metadata – practice E-Periodica

E-Periodica has an OAI data-provider function. The metadata can be obtained via the OAI-PMH interface right down to article level.

An openURL interface enables publishing houses and companies to submit requests with filter options. Articles in PDF format can be accessed directly.

## 5.2. Recording Digitization Projects

The indexing of digitization projects improves their traceability on the one hand and prevents a document from being digitised repeatedly on the other. Projects and contents can also be registered on central, nationwide indexes, e.g.:

- [Digicoord](#)
- [MICHAEL-Portal](#)

- [Zentrales Verzeichnis digitalisierter Drucke \(ZVDD\)](#)
- [Deutsche Digitale Bibliothek](#)

### Practical example – records E-Periodica

The journals at E-Periodica are recorded in different sources:

- [Digicoord](#) (as soon as the digitization of a journal is being planned)
- [ETH-Bibliothek's Knowledge Portal](#)
- [NEBIS](#) and [swissbib](#)
- [EZB](#) – Electronic Journals Library
- [ZDB Zeitschriftendatenbank](#)
- Links and referrals to websites of the publishers and companies involved

## 6. Sources

In the following, selected sources for general works, digitization projects and platforms as well as service providers for hardware and software are being listed.

### 6.1. General Works

- [BCR's CDP Digital Imaging Best Practices Version 2.0](#) (2008) Bibliographical Center for Research
- [Bestandsaufnahme zur Digitalisierung von Kulturgut und Handlungsfelder](#) (2007) Fraunhofer Institute for Intelligent Analysis and Information Systems (pub.)
- [DFG-Praxisregeln "Digitalisierung"](#) (2013) German Research Foundation (DFG) (pub.)
- [Farbmanagement](#) (2014) Spangenberg Theo, Krüger Daniela
- [Good practice guide for developers of cultural heritage web services](#) (2015) UKOLN, University of Bath
- [Good practices in digitization](#) (2007) MINERVA EC (Ministerial NEtwork for Valorising Activities in digitization, eContentplus - Supporting the European Digital Library)
- [Rechtliche Rahmenbedingungen für Digitalisierungsprojekte von Gedächtnisinstitutionen](#) (2015) Weitzmann John H., Klimpel Paul

## 6.2. Digitization Projects and Platforms

### ETH-Bibliothek – collaborations

- [Archives of Contemporary History](#) – Archival material from the Archives of Contemporary History (AfZ) at ETH Zurich
- [e-manuscripta.ch](#) – Manuscript sources from Swiss libraries and archives
- [ETH E-Pics](#) – Photographs, images, illustrations
- [e-rara.ch](#) – Old prints from Swiss libraries
- [E-Periodica](#) – Journals from Switzerland

### Projects all over Switzerland

- [Codices Electronici Sangallenses](#) – CESH, medieval codices of the Abbey Library of St. Gallen
- [Collection of Swiss Law Sources](#) – digitization by the Swiss Law Sources Foundation
- [DigiBern](#) – Bernese culture and history on the internet, Universitätsbibliothek Bern
- [Digicoord](#) – overview over digitization projects in Swiss libraries
- [Dodis](#) – database on Diplomatic Documents of Switzerland, DDS, Berne
- [e-codices](#) – digitised manuscripts from Switzerland
- [Griechischer Geist aus Basler Pressen](#) – catalogue of early Greek prints from Basel in texts and images, Universitätsbibliothek Basel
- [Journal de Genève](#) – cooperative digitization project between the Swiss National Library, the newspaper *Le Temps* and the Bibliothèque de Genève

### Other projects and platforms

- [Bavarikon](#) – Kultur- und Wissensschätze Bayerns, Bayerische Staatsbibliothek, Munich
- [Compact Memory](#) – internet archive for Jewish periodicals, Frankfurt am Main
- [DigiZeitschriften](#) – various fields, DigiZeitschriften e.V., SUB Göttingen
- [Gallica](#) – French literature from various fields, Bibliothèque nationale de France, Paris
- [Göttinger Digitalisierungszentrum](#) – various fields, SUB Göttingen
- [JSTOR](#) – various fields, fee-based, JSTOR, New York
- [Project Euclid](#) – contents from mathematics, predominantly fee-based, Cornell University Library, Ithaca & Duke University Press, Durham
- [Research Library for the History of Education](#) – educational journals, German Institute for International Educational Research, Berlin
- [Zeitschriften der Aufklärung](#) – 18<sup>th</sup> to early 19<sup>th</sup> centuries, Universitätsbibliothek Bielefeld



## 6.3. Service-Providers

### Scanning service-providers

- [4Digitalbooks-Assy SA](#), Ecublens
- [Alos Schweiz](#), Obfelden
- [Bürgerspital Basel](#), Mikrografie, Basel
- [dreischiibe](#), Digital- und Printmedientechnik, St. Gallen,
- [Fachlabor Gubler AG](#), Felben-Wellhausen
- [Herrmann und Kraemer GmbH & Co.KG](#), Garmisch-Partenkirchen
- [digital humanities lab](#), Basel
- [Satz-Rechen-Zentrum \(SRZ\)](#), Berlin
- [Secur'Archiv](#), Carouge
- [SUPAG Spichtig und Partner AG](#), Dällikon
- [Tecnocor ACC AG](#), Kriens

### Scanner hardware

#### *Reflected light scanners / document scanners*

- [CRUSE GmbH Digital Imaging Equipment](#), Wachtberg.  
Sales: [Walter Nagel GmbH & Co. KG](#), Bielefeld
- [i2S SA, Parc Technologique Europarc](#), Pessac
- [Kodak Alaris Germany GmbH](#), Stuttgart
- [MICROBOX GmbH](#), Bad Nauheim
- [Janich & Klass Computertechnik GmbH](#), Waiblingen
- [Zeuschel GmbH](#), Tübingen-Hirschau
- 

#### *Scanning robots*

- [4DigitalBooks – ASSY SA](#), Ecublens
- [Kirtas Technologies, Inc.](#), Victor (NY)
- [Qidenus Technologies GmbH](#), Wien
- [TREVENTUS Mechatronics GmbH](#), Wien

#### *Special scanners*

- [Atiz Innovation](#), Bangkok  
Sales: [Walter Nagel GmbH & Co. KG](#), Bielefeld
- [Grazer Buchtisch](#)
- [Wolfenbütteler Buchspiegel](#)

- [ScanBull Software GmbH](#), Hameln

## Software

### *Workflow, presentation*

- Agora: [Satz-Rechen-Zentrum \(SRZ\)](#), Berlin
- [CCS](#): Content Conversion Specialists GmbH, Hamburg
- [Goobj](#): c/o Sächsische Landesbibliothek, Dresden
- [ImageWareComponents GmbH](#), Bonn
- [Limb Software](#): i2S , Pessac
- Olive: [Olive Software, Inc.](#), Aurora (CO)
- [Visual Library](#): manufacturer semantics, Kommunikationsmanagement GmbH, Aachen  
sales: [Walter Nagel GmbH & Co. KG](#), Bielefeld

### *OCR – Optical Character Recognition*

- [Abby Finereader](#): ABBY Europe GmbH, München

### *Image processing*

- [Adobe Photoshop](#): Adobe Systems Software Ireland Limited, Dublin
- [Adobe Bridge](#): Adobe Systems Software Ireland Limited, Dublin
- [ImageMagick](#): ImageMagick Studio LLC
- [Irfanview](#): Irfan Skiljan
- [PageImprove](#): 4DigitalBooks – ASSY SA, Ecublens

### *File designations and statistics*

- [JoeRenamer](#): Wirth IT-Design, Munich
- [AWStats](#)