DFG Practical Guidelines on Digitisation

UPDATED VERSION 2022

Editors

Reinhard Altenhöner Berlin State Library Prussian Cultural Heritage
Andreas Berger Historical Archives of the City of Cologne
Christian Bracht German Documentation Centre for Art History – Bildarchiv Foto Marburg
Paul Klimpel iRights.Law
Sebastian Meyer Saxon State Library - Dresden State and University Library
Andreas Neuburger State Archives of Baden-Württemberg
Thomas Stäcker University and State Library Darmstadt
Regine Stein Göttingen State and University Library

Contributors

Eva Bodenschatz Saxon State Library - Dresden State and University Library
Barbara Fichtl Göttingen State and University Library
Oliver Götze State Archives of Baden-Württemberg
Melanie Gruß Institute for Theatre Studies, University of Leipzig
Angela Kailus German Documentation Centre for Art History – Bildarchiv Foto Marburg
Celia Krause German Documentation Centre for Art History – Bildarchiv Foto Marburg
Clemens Neudecker Berlin State Library Prussian Cultural Heritage
Stefanie Rühle Göttingen State and University Library
Thomas Scheidt German Documentation Centre for Art History – Bildarchiv Foto Marburg
Zoe Schubert Berlin State Library Prussian Cultural Heritage

DOI 10.5281/zenodo.7561148

Citation: Altenhöner, Reinhard, Berger, Andreas, Bracht, Christian, Klimpel, Paul, Meyer, Sebastian, Neuburger, Andreas, Stäcker, Thomas, & Stein, Regine. (2023).
DFG Practical Guidelines on Digitisation. Updated version 2022. Translated by Eva Bodenschatz. Zenodo.
https://doi.org/10.5281/zenodo.7561148

German version: https://doi.org/10.5281/zenodo.7435724

Contents

Introduction

In its "Digitisation and Indexing" programme, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)[1] funds projects that expand and improve availability of digitised objects and metadata for research. The aim of the programme, which has extended to all scientifically relevant object types after being opened up in 2020, is to digitise and/or index holdings and collections that are of transregional importance for research. In addition, it promotes the development of material-specific quality criteria and practical guidelines, where these do not yet exist.[2]

The Practical Guidelines on Digitisation provide a fundamental basis for DFG-funded digitisation projects: They are intended to make it easier for applicants to plan digitisation projects and to support a comparative review of proposals. The Practical Guidelines formulate standards and contain information on organisational, methodological and technical issues in the context of digitising and indexing objects relevant for research. They thus make an important contribution towards the sustainability, accessibility and compatibility of funded projects and to the resulting infrastructure.

In the field of manuscript and print tradition, practical guidelines for digitisation are available since 2007 and were updated in regular intervals under the coordination of the DFG Head Office. Since 2013, they also include standards and recommendations for digitising archival materials and image-based objects in museum collections. The last revision was made in 2016. Opening up the funding programme, the diversification of the material to be digitised and the high speed of technical advancements have resulted in the need to revise and differentiate the existing Practical Guidelines. The DFG has recognised this need and has placed the further development of the Practical Guidelines on the broader foundation of a self-organised process that is better able to reflect the differentiated needs of the respective communities.[3] In support of this, the DFG provides suitable instruments for funding.[4]

This document is an updated version of the Practical Guidelines last published in 2016. It was developed in consultation with the DFG Head Office by a group of authors initiated by the NFDI consortium NFDI4Culture, the majority of whose members have long played a part in shaping the Practical Guidelines and are actively involved in the NFDI consortia NFDI4Culture, NFDI4Memory, NFDI4Objects and Text+[5].

---

[1] The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) is the central self-governing research funding organisation that promotes research at universities and publicly funded research institutions in Germany. The DFG serves all branches of science and the humanities by funding research projects and by fostering collaboration among researchers (http://www.dfg.de/en). The DFG also supports projects to establish and further develop high-performance information systems and infrastructures for research. The results of funded projects must be accessible to researchers at no charge and for the long term (http://www.dfg.de/lis/en/).

All web links given here and below were last checked on 15 December 2022.

[2] See https://www.dfg.de/en/research_funding/programmes/infrastructure/lis/funding_opportunities/digitisation_cataloguing/index.html.

[3] In order to initiate this process, a DFG roundtable discussion on the self-organisation of the Practical Guidelines on Digitisation was held on 26 April 2021, see https://www.dfg.de/download/pdf/foerderung/programme/lis/rundgespraech_praxisregeln.pdf.

[4] https://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/vigo/.

[5] https://nfdi4culture.de/, https://4memory.de/, https://www.nfdi4objects.net/index.php/en/, https://www.text-plus.org/en/home/.

The previous document structure has been retained while the complete text was reviewed; essential parts of the text remain unchanged and valid. Updates concern some technical aspects, especially in the area of 3D digitisation, legal aspects against the backdrop of the copyright reform in 2021 and the appendices referring to the recommended application profiles for metadata.

Section 1 briefly summarises the most important requirements. Sections 2 to 6 introduce in detail the issues and methods relevant to all organisations seeking to digitise objects. These sections are also geared toward institutions which are about to plan digitisation projects and further address digitisation measures in the absence of detailed indexing data. Section 7 specifies the presentation standards and formats required for text-based objects.

The now revised Practical Guidelines on Digitisation serve as starting point for a material- and community-related differentiation of the Practical Guidelines by the communities. The institutions contributing to this document will engage in its continued development. These are: Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg (German Documentation Centre for Art History – Bildarchiv Foto Marburg), Historisches Archiv der Stadt Köln (Historical Archives of the City of Cologne), Landesarchiv Baden-Württemberg (State Archives of Baden-Württemberg), Niedersächsische Staats- und Universitätsbibliothek Göttingen (Göttingen State and University Library), Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (Saxon State Library - Dresden State and University Library), Staatsbibliothek zu Berlin Preußischer Kulturbesitz (Berlin State Library Prussian Cultural Heritage), Universitäts- und Landesbibliothek Darmstadt (University and State Library Darmstadt). In line with their joint Memorandum of Understanding , the NFDI consortia NFDI4Culture, NFDI4Memory, NFDI4Objects and Text+ are available as points of contact and platform regarding the next steps, as the Practical Guidelines on Digitisation are a central component in forming standards in their research communities. All communities and institutions concerned with the digitisation of research-relevant objects are encouraged to contribute their expertise to the further process.

1. Checklist for Applicants and Reviewers

As explained in the following sections, there are numerous choices to make when planning digitisation projects. Rigidly prescribing specific standards would therefore unduly restrict the projects to be funded and hamper their continued dynamic development. However, the guidelines provide information on risks with respect to project planning and realisation. The checklist for applicants and reviewers should therefore be understood and applied as follows:

- For all proposals with digitisation components, the technical concepts will be reviewed regardless of their content evaluation.
- Reviewers will check whether the necessary rights clearances regarding the source material have been completed and substantiated, as well as the precautions that allow full reusability of project results (→ 1.8).
- Proposals must demonstrate plausibly that the project will be implemented according to the standards listed below.
- Any deviation from these standards needs to be justified.
- For better comparability, proposals (or progress reports) should contain information on estimated (or actual) costs for technical reproduction including structural data capture per digitised item and title / index unit, as well as costs for full text capture.[6]
- If it is intended that a digitisation project should not comply with the standards listed below, the need for doing so should be explained in detail if this entails higher costs.
- The technical preparation of a proposal must be comprehensive enough to allow an overall evaluation of technical requirements and procedures of the proposal. While the initial project stage may include the testing of innovative technology, it cannot be used to determine, for instance, how long the digitisation

[6] See the statistics sheet on new and renewal proposals within the "Digitisation and Indexing" programme https://www.dfg.de/formulare/12_152/12_152_de.xlsx. There are currently no commonly accepted standards for calculating real costs. The Practical Guidelines therefore recommend that the costs of scanning including structural data entries be calculated as follows:

(1) For outsourced digitisation:
   a) Amount of money paid to service provider per digitised item.
   b) Proportionate costs of all flat fees charged by service provider (e.g. for naming and storing files, issuing structural data, DVDs for transferring data from service provider to project location).
   c) Proportionate costs for staff members occupied exclusively, or to a calculable extent, with digitisation quality control.

(2) For in-house digitisation:
   a) Proportionate costs of newly acquired digitisation hardware in the narrower sense. Hardware is considered depreciated when the project ends (actual entire duration; e.g. in the case of an ongoing four-year project that was initially proposed as a two-year project, the base is the projected number of items to be digitised throughout the full four-year period).
   b) Proportionate personnel costs for all personnel who operate digitisation hardware and issue structural data.
   c) Proportionate costs for staff members occupied exclusively, or to a calculable extent, with digitisation quality control.

In both cases, (1) and (2), costs must be calculated per digitised item. Expenses related to the following tasks are not considered costs for scanning:

   a) Project management (e.g. selecting and fetching materials to be digitised).
   b) Entry of metadata other than structural data.
   c) Long-term archiving.
   d) Indirect costs typically assessed in terms of internal cost accounting.

campaign will take, what grade of digital copies should be produced, or how the general workflow should be designed. Any pilot studies necessary to resolve such issues must be concluded prior to the submission of a proposal.

1.1 General technical procedures and prerequisites

The proposal must describe the intended workflow in sufficient detail to allow reviewers to assess the following questions:

- Is the funding for staff both sufficient and necessary? To help answer this question, average available resources (working hours, storage capacity of computers used in the workflow) per unit must be stated.
- Are the projected processing times realistic? To the extent that the projected processing times are not immediately plausible, they should be substantiated either by experiences gained in previously completed, similar projects, by published and recognised benchmarks, or by the results of self-conducted pretests.

1.2 Technical parameters of digital reproduction

The aim of digitisation is to reproduce the original material as faithfully as possible, according to scientific requirements. The following guidelines for minimum digitisation requirements refer only to digital masters (→ 3.2.1) and are general recommendations especially for pictorial and textual material. For material-specific recommendations that exceed the scope of these minimum requirements, see section (→ 3.2.2). Comparable quality standards for digital copies of other material groups are in preparation.

- The resolution should be such that the smallest relevant details are clearly visible in the archive copies when the file is reduced to a quarter of its original size. As a general rule, a resolution of at least 300 ppi is recommended (→ 3.2.1.1).
- For storage of the digital master, a colour depth of 16 bit in RGB colour mode is sufficient (→ 3.2.1.2).
- Masters should be stored as uncompressed Baseline TIFF files. The more advanced options of extended TIFFs should not be used for digital masters. In addition to TIFF, TIFF-LZW or lossless JPEG2000 may also be used as an image master format. In general, however, formats without compression are preferable to those with lossless compression. When storing masters in JPEG2000 format it is important to ensure that only the public-domain parts of JPEG2000 are used. For the risks associated with JPEG2000 and TIFF-LZW with regard to long-term archiving, see section → 3.2.1.4.
- In the spirit of the desired open and free access to any form of scientific information (open access), it is expected that digital reproductions are offered as freely re-usable as legally possible.

1.3 Metadata

Metadata should be provided in a software-neutral, standards-compliant form, generally XML encoding.

- To ensure the optimally distributed and long-term usability of descriptive metadata, indexing should be based on relevant standards and reference models and, wherever possible, linked with published authority data. To record personal, biographical and geographic information, the use of the Integrated

Authority File (GND)[7] offered by the German National Library is expected. Other controlled vocabularies such as Iconclass for image classification must be nationally and internationally compatible (→ 3.3.1).

- The provision of descriptive metadata for further use in accordance with material-specific standards is mandatory. The DFG Basic Data Record[8] can be used for minimal indexing; for further recording, the creation of a core field catalogue is recommended (→ 3.3.1).

For metadata to describe the object in portals content-wise in one of the following formats:

- Text Encoding Initiative (TEI[9]) for textual sources
- Lightweight Information Describing Objects (LIDO[10]) for pictorial and three-dimensional objects
- Encoded Archival Description (EAD[11]), common application profile: EAD(DDB)[12]

For metadata to be displayed in the DFG Viewer (see below) in one of the following formats:

- Metadata Encoding and Transmission Standard / Metadata Object Description Schema (METS[13]/MODS) for printed text-based works and archive material
- Metadata Encoding and Transmission Standard / Text Encoding Initiative (METS/TEI) for manuscripts

The metadata must be valid against the relevant XML schema and must also be checked for semantic correctness.

- Persistent identifiers must link the digital copies and metadata in the reference system (catalogue, online search system). Metadata capture and digitisation must be coordinated with or performed by a scientific infrastructure facility. Digitised items should be listed in relevant portals and network systems. All materials must be presented in suitable subject-specific or cross-disciplinary online applications (→ 3.3).
- Data should be delivered to these portals in accordance with standard formats, if possible via an OAI-PMH[14] interface. All institutions and target portals that do not offer an OAI interface themselves can also specify conventional http URIs as a direct link to the METS files. The Deutsche Digitale Bibliothek (DDB[15], German Digital Library) and Archivportal-D do not yet offer a publicly accessible OAI interface for data deliveries (→ 3.3.4).
- It is the responsibility of the grant recipient to ensure that the digital copies produced in the project (images as well as full text) can be unambiguously identified, and searched and retrieved separately from other units recorded in the same system (→ 3.3).
- The decision whether to generate structural metadata only concerns text documents. If structural metadata is used, it is recommended to consult the structural data list available on the DFG Viewer[16] website. The standards currently recommended are METS or TEI. However, the DFG Viewer should be chosen for text-

[7] https://gnd.network/.
[8] https://www.dfg.de/formulare/12_155/.
[9] https://tei-c.org/.
[10] https://cidoc.mini.icom.museum/working-groups/lido/lido-overview/.
[11] https://www.loc.gov/ead/index.html.
[12] https://wiki.deutsche-digitale-bibliothek.de/x/hBIiAQ.
[13] http://www.loc.gov/standards/mets.
[14] http://www.openarchives.org/.
[15] https://www.deutsche-digitale-bibliothek.de/.
[16] http://dfg-viewer.de. In the future, the Practical Guidelines should no longer specify certain viewer technologies.

based materials unless there are content-related reasons against it. Solutions for presenting image material may be based on the International Image Interoperability Framework (IIIF[17]) (→ 3.3.2).

- As a minimum, digitisation projects should present the nature and scope of the selection of objects on a website, preferably with an English-language version. Standardised collection descriptions are desirable, typically in the same format as the object descriptions: MODS or TEI-Header, EAD(DDB), LIDO (→ 3.3.3).[18]

1.4 Full text generation

All proposals to digitise text-based objects are expected to address the option of full text provision. For printed works dating from 1850 or later, the full text must be generated and mere image digitisation is not sufficient. (→ 3.4). Full text includes the characters of the master copy, markup data to identify structural features, and metadata, which usually are part of the same file (→ 3.4).

- Full text can be generated in two ways: Through Optical Character Recognition/Handwritten Text Recognition (OCR/HTR) or transcription (→ 3.4.1).
- To create some uniformity on which to base the evaluation of accuracy, applicants are requested to state the letter accuracy, i.e. incorrect syllabification and layout errors should be ignored. Random samples should be based on the statistical procedure described in section 3.4.1. (→ 3.4.1).
- For the purposes of character encoding, it is recommended to save the text in Unicode. UTF-8 should be preferred (→ 3.4.2).
- Unless there are cogent reasons not to, full texts must be encoded and marked up. For descriptive markup and structured full text the Text Encoding Initiative (TEI) model is generally recommended; for layout information and coordinates, ALTO XML (ALTO V4.x[19]) and PAGE XML[20] have proved their worth alongside TEI (→ 3.4.3).
- In some cases, it is important for the presentation of full text to preserve the layout of a document for the long term. The Practical Guidelines recommend using a suitable formatting language (e.g. XSLT, XSL-FO, XQuery or CSS), which largely ensures independence from special software. If valid reasons prohibit archiving the format with XML techniques, layout information for text documents may also be archived in PDF according to ISO standard 19005-1. However, as explained in section 3.4.3., PDF files cannot replace the provision of marked-up full text in XML (→ 3.4.4).
- In keeping with the principles of open access and open source, it is expected that machine-readable full texts or the XML on which these texts are based, XSLT scripts and DTDs or XML schema files will be made available for reuse as freely as legally possible (→ 1.8; 6.1).

1.5 Long-term availability

At present, there is no universal solution for the long-term preservation and archiving of digital content that is suitable for all types of objects. For long-term preservation, files are stored in stable, migratable formats in a technically and organisationally secure storage system.

---

[17] https://iiif.io/.

[18] The RDF-based Data Catalog Vocabulary (DCAT) is also becoming increasingly relevant: https://www.w3.org/TR/vocab-dcat-3/. See also Data Catalog Vocabulary (DCAT) - Version 2. W3C Candidate Recommendation 04 February 2020: https://www.w3.org/TR/vocab-dcat-2/.

[19] https://www.loc.gov/standards/alto/. Since 2018: https://github.com/altoxml/documentation/wiki/Versions.

[20] https://ocr-d.de/de/gt-guidelines/trans/trPage.

Digital data is archived with a similar storage system but with more extensive technical and organisational measures that cover not only the physical retention of the data but also strategies to make them available for use (access). OAIS[21] (Open Archival Information System) serves as reference model (→ 3.5).

The long-term availability of the results of digitisation projects depends firstly on the choice of data and metadata formats. Secondly, it must be ensured that the digital data remains physically available. It should be noted that in DFG-funded digitisation projects, the costs of project-specific data preservation are expected to be borne by the institution for the duration of the project. DFG funding is not available to cover these costs (→ 3.5).

It should be noted that the DFG views digitisation projects as endeavours by the entire institution: It is assumed that the department carrying out the project will be supported by the in-house IT infrastructure. We also encourage smaller institutions to take advantage of the expertise and services of larger institutions (→ 3.5).

Proposals must contain convincing statements as to institutional long-term preservation and archiving. Digitised material should be archived redundantly (→ 3.5).

1.6 Organisational issues, collaboration with service providers

The decision whether to undertake digitisation as an in-house project or to outsource it is always specific to the project and exclusively the applicant‘s responsibility. Services must be precisely defined by contract. The DFG expects that an appropriate percentage of the invoice amount will be retained for security purposes and will only be paid to the service provider after a quality check has been performed (→ 4).

1.7 Citing, persistent addressing

Digital representations must have a unique address so that other objects or databases can link to them. In addition to the customary citation format, which can and should still be given via the navigation software, this requires the specification and online documentation of addressing techniques.

The accessibility and citability of the resource as a whole and the work’s individual physical parts must be guaranteed. Institutions should implement suitable mechanisms (Persistent Uniform Resource Locator (PURL), Uniform Resource Name (URN), Digital Object Identifier (DOI), Handle, etc.) to ensure the persistence and linkability of a resource, thus reliably providing sources for scientific research. For printed text documents, it is recommended to generate URNs via the German National Library (→ 5).

1.8 Provision of digital resources and metadata to the public

The DFG funds the digitisation of research-relevant material in order to make this material available to researchers in Germany and around the world. The following aspects should be considered:

Rights relating to the objects to be digitised must be cleared before the project begins. In particular, any copyrights, personality rights, ancillary rights and archive rights must be taken into account (→ 2.2 and 6.1).[22]

[21] See: http://d-nb.info/104761314X/34.

[22] A rights clearance is not always possible prior to the start of a project. In justified exceptional cases, a rights clearance is permitted within the framework of the project (possibly even with funding).

Metadata, digital copies, full texts and XML files should be made available for reuse as freely as legally possible. They must be marked as public domain or be offered under the most liberal Creative Commons licence[23] possible. Standardised rights notices must be used for protected material (→ 6.1).

Digital masters should be provided as high-resolution derivatives in TIFF format or full-resolution JPEGs. For AV material, the minimum standard for presentation derivatives should be MP3 (audio) or MP4 (video), and glTF/GLB for 3D media. The data should be available without technical access restrictions, such as login, IP address, etc. (→ 6.1).

For projects in which not only public domain material is digitised, delayed open-access publication (a moving wall) of up to one year after the end of the project may be agreed upon (→ 6.1).

A restriction on open access may be legally justified, but it may not affect more than 10% of the total material to be digitised and requires submission of a plan for the timeline and technical implementation of material availability.[24]

All objects are to be provided with URLs with the lowest possible granularity and which are citable in the long term. The form in which citations are to be made must be clearly indicated.

The DFG expects that data made available online as part of DFG-funded projects should include a clear reference to its origin and if applicable a reference to the DFG funding (→ 6.1). Accessibility should be realised in different ways (→ 6.2.1)

- via the website of the institution carrying out the project
- via the local or regional catalogue
- via national reference and presentation systems, in particular the DDB and the Archivportal-D
- via material-specific portals (if available)
- via interfaces
- All currently popular browsers must be supported to the extent that this is objectively viable (→ 6.2.2 and 7.2).

2. Objectives and Selection

2.1 Objectives

Digitisation has become a vital research tool in the humanities, cultural studies and the emerging field of digital humanities. It has made direct research with sources much easier, while conserving valuable and sometimes fragile originals. Not only does the digitisation of library, archive and museum holdings make copies easily accessible online, it also helps build an infrastructure that turns the Internet into an integral research space for the increasingly digital world of research in the humanities and cultural studies. Only by linking these digital objects with other online resources can the potential of the Internet be fully leveraged. Thus, the objective is not only to make these materials available and usable, but also and especially to interconnect the different resources to form a virtual research infrastructure.

[23] See licences in the currently valid version: https://creativecommons.org/.

[24] The DFG cannot fund the processing of non-open-access material.

While there is verifiable knowledge for implementing digitisation projects, these insights must not be applied mechanically: What constitutes essential conservational care when digitising medieval manuscripts may be unnecessarily time-consuming and expensive when mass processing administrative records from the late 19th century.

Since the majority of DFG-funded projects to date have focused on flatware (2D), the techniques and parameters pertaining to this area will be explained in detail below. Besides textual material, image-based material is significantly gaining in importance. Therefore, this edition of the Practical Guidelines continues to integrate developments in this area, especially concerning the technical production, without claiming to be exhaustive. With regard to formats and digitisation methods, the guideline also deals with AV media or three-dimensional collection objects in certain points, but still lacking the necessary depth.

2.2 Selection

In general, it should be noted that the technical aspects of digitisation can be planned quite well, while the intellectual effort required to select the right items is difficult to calculate. A decision must therefore be made in each case as to whether the inclusion of a greater number of documents will ultimately be cheaper and more efficient than undertaking a complex evaluation and selection process. It is highly recommended to take advantage of existing selections and reference works such as bibliographies and subject databases. The basic selection criteria are scientific relevance and demand by researchers.

Defining a corpus under the criteria of scientific relevance or demand by researchers is not always easy. In difficult situations, the case for a project may be made by cooperating with a specific research community or institution, which can plausibly formulate its own needs. Successful projects are often based on cooperative arrangements in which an academic undertaking, e.g. a research or editorial project in philology or legal history, seeks to establish an online presence and links back to library or archive holdings, thus enabling two-way linkage.[25]

Concepts of digitisation on demand and digitisation on use (the digitisation of commonly or frequently used material in libraries and archives) are certainly also interesting, as they ensure a concrete interest in use.

In general, more specialised digitisation projects that focus on a specific research question should provide interfaces to deliver their data to larger projects with a more formal structure. Conversely, interdisciplinary projects should open up their results to more in-depth and specialised reuse.

The DFG funds the digitisation of research-relevant material in order to make this material available to researchers in Germany and around the world. Clearance of rights to the material must be completed when project planning begins, or at the latest when the proposal is submitted, and must be substantiated in the proposal. In particular, any copyrights, personality rights, ancillary rights and archive rights must be taken into account. Rights clearance often proves to be more time-consuming than anticipated. Sufficient time and personnel should therefore be allocated to this task. If the clearance of rights proves to be difficult, we recommend seeking professional support, for example from the institution's own legal department. Even during the selection process, it should be noted that the results of DFG-funded digitisation projects must be made available to the research community as freely as possible by way of open access. Wherever material cannot be assigned to the public domain, a free Creative Commons licence must be chosen (CC0, CC-BY or CC-BY-SA) (→ 6.1). Only in justified exceptional cases is online

[25] Research projects are generally based on funded infrastructures. However, it is also possible to apply for funding for projects which combine both research and infrastructure elements. Alternatively, relevant specialised bibliographies that formulate a canon, or to a mixture of both, may be used.

publication also permissible under the statutory permissions (exceptions) for orphan works (§§ 61 et seq. UrhG (German Copyright Act)) or unavailable works (§§ 61d et seq. UrhG) as these do not meet the requirement for an open access publication.

2.3 Published texts: duplicate checking and data matching

To avoid redundant digitisation of printed textual sources it is sensible to check, before submitting a proposal, whether the works selected for digitisation are already digitally available in Germany or elsewhere. The following requirements apply:

Proposals and reports are expected to mention finalised or ongoing national and international digitisation projects to the extent that they relate to the proposed or ongoing DFG project and the materials it covers.

For large-scope projects (e.g. over 1,000 printed works), proposals and reports should explain how they relate to commercial digitisation offerings that are accessible free of charge. A pragmatic effort should be made to keep the number of duplicate digitisations as small as possible.

For printed works published between 1501 and 1800, the VD16, VD17 and VD18 bibliographies should be consulted to check for existing digitisations. URNs and PURLs of digital copies must be reported to these bibliographies. For digitisations of incunabula, the Census of Incunabula for Germany (ISTC) should be consulted.

3. Digitisation and Indexing

Retrodigitisation in the area of 2D always includes creating digital derivatives (of the digital master → 3.2.1) and producing metadata, and, in the case of text-based works, it may additionally involve capturing full text and generating structured data and mark-up. Digital images also refer to genuine digitally created (born digital) reproductions of material objects (e.g. digital photographs of buildings). In the following recommendations, the term digitisation refers to the entire workflow (preparation, digitisation proper, the generation of bibliographical metadata, structural data, full text, and long-term safeguarding / digital preservation). No detailed distinction will be made between different material types (e.g. text-based works, pictorial works such as paintings, drawings, printmaking, film and audio material such as video, DVD, CD, vinyl records, music cassettes as well as commodities, stationary objects / buildings etc.) in the following.

3.1 Preparation of materials and conservation review

Preparatory activities in digitisation projects are often underestimated and should be carefully taken into account when planning a project. Are the objects actually available? Are there any conservation-related objections to digitising the originals? Are there sufficient personnel to draw out and prepare the objects? Are employees with academic or bibliographic training available to perform completeness checks or collations, if catalogue entries do not provide this information? In the case of printed text works, digitisation of incomplete or defective samples should be avoided where possible and reproduction of a complete sample should be strived for. This equally applies to other types of material.

Although the conservation review may be very time-consuming, it should not be omitted under any circumstances in order to protect the objects. We recommend using checklists to assess suitability for digitisation and drawing up

a set of digitisation guidelines on this basis.[26] If reproduction could expose an original to risk or undue stress,[27] it should be done from an existing microfilm[28], if possible, or be avoided altogether. At any rate, valuable historic materials must be handled with due conservational care, even if this reduces the digitisation throughput and takes more time. Scanning systems and equipment such as book supports or securing aids must comply with the guidelines of the conservation review.

3.2 Technical parameters of digital reproduction

The aim of digitisation is to reproduce the original object as faithfully as possible, according to scientific requirements. Digitisation parameters should be selected with regard to the quality of the desired reproduction, long-term availability, and interoperability.

3.2.1 General explanations and parameters (two-dimensional objects)

Two types of reproductions are important in digitisation: The digital master, i.e. the source or archive format, and derivatives generated for users. The digital master forms the basis for all further processes. Its production should therefore be given special attention and the relevant recommendations should be taken into account with regard to long-term storage. The following guidelines for minimum digitisation requirements refer only to digital masters.

3.2.1.1 Resolution

The resolution of a file is normally measured in ppi (pixel per inch)[29]. ppi as a ratio always relates to a page dimension. The total dimension of an image file is the pixel dimension.

The minimum scan resolution chosen for digital copies should allow the details of the original to be fully reproduced at life size.[30]

---

[26] A conservation review should include the following criteria: Risk to ink; risk to paint layer / priming coat; ink / paint corrosion; loose stitching; binding or cover material too stiff / inflexible; binding (partially) fractured; cover material at joint (partially) fractured; spine inlay too stiff; spine binding damaged; cracks, tears or flaws; delicate leather grain on spine; risk to spine gilding; very thick vellum volume with hollow spine; cover (partially) fractured; cover loose (front/back); closure straps too stiff / partially fractured, microbial damage; microform available. There are also some technical problems that make digitisation difficult or impossible: Inside margin too narrow; book block corrugated; leaves / signatures very stiff; unavoidable loss of text; original cannot be laid flat; extreme format.

[27] Corbach, Almuth: Bestandsschonendes Digitalisieren von schriftlichem Kulturgut. In: Digital und analog. Die beiden Archiv-welten. 46. Rheinischer Archivtag. Ratingen 21.-22. Juni 2012. Beiträge. Ed.: LVR-Archivberatungs- und Fortbildungszentrum. Bonn 2013 (Archivhefte 43), pp. 90 – 102.

[28] In archives, it is standard practice to digitise microfilm regardless of conservational considerations. For a discussion of the material-specific parameters affecting the digitisation of microfilm, see section (→ 3.2.2.4).

[29] In contrast to printing, ppi (pixel-per-inch) has now become the standard for digital image data.

[30] One indicator of the required minimum resolution for a digital copy is the resolution of the human eye at a comfortable reading distance. The shortest distance at which a person can observe an object for long periods without fatigue is about 25 cm. The human eye is able to distinguish two lines when light falls on at least two non-adjacent photoreceptors with at least one photoreceptor between them. At a viewing distance of 20 cm, a raster frequency of 60 lines/cm can no longer be perceived as separate lines. This has produced the following resolution requirement, which is also standard in the printing industry:

A raster of 60 lines/cm requires at least 120 pixels.
120 px/cm x 2.54 cm/in = 304.8 ppi, or approximately 300 ppi

The basic recommendation for minimum target resolution is therefore 300 ppi relative to the format of the original (output format of digital copy at 300 ppi). However, 300 ppi only relates to originals intended to be viewed with the naked eye, such as text-based works, graphic works and photographs. Different guidelines apply to media whose complete visual information is only visible when enlarged. This includes miniatures of any kind and especially photographic transparencies (e.g. negatives or slides).[31]

There are also objects not covered by the procedure to calculate resolution, as described above, due to their dimensions, in particular works of architecture, but also oversized paintings. For instance, a resolution of 300 ppi with a DIN A2 output format can be achieved with a 50 megapixel digital camera.[32]

Guidelines on advisable deviations from the basic recommended resolution of 300 ppi relative to the original format are given along with the material-specific parameters (→ 3.2.2).

Simply the number of available pixels in a scanning line or sensor does not govern the resolution that can be achieved with a digital capture system. It is dependent on various factors, for example the technical quality of the digital recording device (camera or scanner) and the reproduction capability of the lenses used.

3.2.1.2 Colour depth

Colour depth determines the differences between brightness and colour values in a digital image. Because digital technology only uses discrete statuses (yes/no), unlike analogue photography, it cannot represent brightness and colour differences continuously with fluid transitions.

The advantage of the higher colour depth, of 16 bits per channel, lies in the greater colour differentiation.[33] It means that fewer tone values are lost during subsequent image processing. This is important in the case of digital images that require a lot of processing, for example black and white negatives.

Digital capture systems generally have a high colour depth of 14 or 16 bits per channel.

3.2.1.3 Digital capture process

Capturing technique

There are two main techniques in photographic digital capturing. The first is line scanning, where a trilinear scan line, of a fixed width, scans a defined section, with one line each for red, green and blue. Flatbed scanners and line scan back parts for specialist or medium-format cameras use this technique. Thereby, the colour information is physically generated for each pixel.

---

[31] Example: Small-format negatives of 24 x 36 mm have been produced for enlargement on photographic paper. They are not suitable for viewing with the naked eye. The general practical guideline for photographs is that small-format negatives (depending on lens, lighting, film and developer) can be enlarged about 10x, in this case producing a photograph of 24 x 36 cm. To achieve these target dimensions of 24 x 36 cm as described above in the digital copy of the negative, the negative must be digitised with at least 2835 x 4252 pixel.

[32] According to the principle that camera pixels / target resolution in ppi equals target dimensions in inches, an area sensor of, for example, 5700 x 8600 pixels yields the following output dimensions at 300 ppi:

8600/300 = 28,7 in x 2.54 cm/in = 72.81 cm

5700/300 = 19 in x 2.54 cm/in = 48.26 cm

[33] From a technical point of view, it should be noted that common image processing programmes (e.g. OpenCV, Pillow, ImageMagick) do not provide full support for 48-bit colour and thus require appropriate conversions (potential source of error) for processing.

The second common technique is the area sensor, where a defined area is equipped with a certain number of photodiodes.

Line scan systems and multi-shot-capable area sensor cameras offer advantages for delicate subjects such as intaglio illustrations, maps and textiles.

Standard flatbed scanners can digitise two-dimensional media up to an original size of DIN A3. However, the process is not contact-free as the medium comes into contact with the scanner's glass pane. For this reason, it is only suitable for originals to which no conservation concerns apply.

Image noise

To reproduce as much detail as possible, the manufacturer's recommendations for reducing image noise should be observed for the chosen camera technology. The following rule of thumb applies: The lower the ISO setting, the lower the noise.[34]

Lenses

We also recommend the use of modern high-quality lenses designed for the high-resolution potential of digital cameras. Prime lenses are to be preferred over zoom lenses, as they allow imaging errors to be corrected more easily and overhead use will damage the mechanics over time. Zoom lenses always deliver a quality compromise over the available focal length range. Shift lenses reduce losses in image quality caused by the later rectification of falling perspective lines. If the reproduction scale is less than 1:10, special macro lenses designed for this type of imaging should be used. Subject to the camera system being used, high-quality magnifying lenses may also be used for reproductions on this scale.[35]

Working area

To prevent blurring, caused by shake, during the digitisation process, digitisation units should be positioned in a location that is not subject to vibration. Wooden floors, for example, are unsuitable because they transmit vibrations easily. If reflex cameras are used, they should be operated with mirror lock-up to prevent vibration when the mirror flips up. Tripods and repro stands should be of a suitable size to support the weight of the cameras easily.

Moiré

In the case of objects with fine, evenly distributed details, using the one-shot technique with a digital camera and area sensor may lead to colour distortion and a moiré effect.[36]

A moiré effect is created in digital imaging when the uniform pixel matrix of a digitally generated image interferes with a regular line structure on the original. This is due to the pixel size of the scan line or area sensor, the

[34] ISO 19264-12021, page 19.

[35] The reproduction capability of a lens can also be identified from the corresponding modulation transfer function (MTF) diagrams. Various lens manufacturers publish these diagrams. The MTF describes how well a lens can transfer the edge contrast, on an original item, to the photographic image. It thus provides an indication of the resolution capability of the lens. The technical data sheets for the lens also provide information about the brightness and sharpness reduction at the edge of the image. Optical imaging errors are especially noticeable here. To reduce image errors a lens should be stopped down twice to exploit the better reproduction capability of the centre. If depth of focus is not important, for example when photographing flat materials, this optimum working aperture should be used.

[36] The moiré discussed here is not to be confused with the phenomenon created by the interference of print dots on the rasterised original with the pixels on the monitor.

reproduction scale, and the line frequency of the original. It is difficult to predict the occurrence of this effect, as the parameters involved rarely remain constant.

Colour distortion occurs with one-shot area sensors under the same conditions but is the result of an interpolation error by the camera software. In a one-shot camera system, each pixel is only sensitive to one primary colour: red, green or blue. The remaining colour information for the pixel is interpolated from the colour information for the surrounding pixels. If the colour and brightness differences in the original item are too extreme, colour distortion will result. In cases like these, line scanners or multi-shot digital backs may yield better results as they physically generate the colour information instead of calculating it. However, this does make digitisation more time-consuming as scanning times are longer and absolutely constant lighting is essential during the capture process.

To avoid the moiré effect, a test digitisation should be carried out to identify the most suitable process and required resolution before carrying out digitisation proper.

Objects with which moiré may occur include printed half-tone documents, copperplate engravings and textiles.

Lighting

The most appropriate lighting should be carefully considered before beginning a digitisation project, with due regard to conservational considerations. Line scan systems require continuous, flicker-free light because the original is continuously scanned for a certain period. The light power and colour must remain constant during scanning to achieve uniform results. With area sensors, a flashgun may also be used. In the case of multi-shot systems, the flash power and colour must remain constant from one exposure to another. The same applies to continuous light.

From a conservation point of view, it can be argued that brief exposure to intense light is less harmful than extreme heat, particularly as it is the overall light exposure that counts, i.e. it does not matter whether the same amount of light is distributed over a longer amount of time or falls on the item for a short period. However, strong short-term fluctuations in brightness are problematic for the staff from an occupational-health point-of-view.

Image processing

Photographs are normally generated in a manufacturer-dependent raw image format (RAW), at maximum size with a colour depth of 14 or 16 bits per colour channel. RAW images represent the original camera data. Image corrections should be undertaken in RAW software if possible. As this file is not changed, it can be corrected again if necessary without loss. Editing rules can also be applied to any number of images as a pre-set, which significantly increases efficiency.

The RAW format as a proprietary file format is not yet suitable for archiving. Nevertheless, it should definitely be preserved, as it always retains the original image information, unlike the edited master TIFF. The current recommendation is to store the master file in the container format DNG.

It is possible to determine whether a digital capture system is suitable for a digitisation project by carrying out a test digitisation with standardised test charts (e.g. ISO 12233 test chart, USAF 1951 test chart). The effective resolution of the test images can then be calculated with the help of analysis software or resolution tables, which are often supplied with the chart. It is recommended to work with colour profiles (input profiles, output profiles) stored in the recording device.

Adobe RGB, the ECI-RGB v2 colour space recommended by the European Color Initiative (ECI)[37], or the identical colour space L-Star RGB[38] should be selected as the output profile for colour images.

To allow for retraceable colour reproduction, it is necessary to include a suitable colour chart or colour wedge during image capture. If reproduction conditions remain the same, the colour reference image can be placed representatively in front of a series.

For brightness and tone value control, it is important not to crop the tone value histogram (for black and white) at the sides so as to retain the full range of tone values. The recommended limits for the reproduction of dark and light tonal values are, according to the Lab colour model:

Black: L*5

White: L*95

Sharpening should only be done moderately for digital masters. Additional sharpening may be used for different derivatives depending on the application.

For reliable matching of originals with the on-screen representation of the digital image, originals should be viewed under D50 standard lighting for graphic workstations in accordance with ISO 3664:2000 (which corresponds to our ideal recommendation). The types of monitors used in the graphic industry provide the most accurate representations in these circumstances.

Digital image postprocessing

Additional post-processing may be required to optimise the quality of digitally captured images. For the digital master, this should be limited to the necessary correction of colour and tone values. Every effort should be made to avoid object deformations, adding or deleting parts of an object, and special effects such as the use of modifying filters. To increase image integrity, positioning aids or backgrounds in derivatives may be removed from the image at a later point.

3.2.1.4 File formats

According to current knowledge, image masters should be archived in TIFF uncompressed[39]. The TIFF format has been around since the 1980s. It has established itself as one of the most important de-facto standards, and it is expected that all standard programmes will continue to support this format. However, this holds true only for so-called baseline TIFFs. The more advanced options of extended TIFFs should not be used for digital masters.

In addition to TIFF, TIFF-LZW or JPEG2000[40] in its lossless form may also be used as an image master format. However, to store masters in JPEG2000 format it is important to note that only the public-domain parts of JPEG2000 may be used.

[37] European Color Initiative: http://www.eci.org.

[38] http://www.colormanagement.org/de/workingspaces.html.

[39] https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=686; http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1099.

[40] http://www.jpeg.org/jpeg2000/.

In the last few years, the ISO standard JPEG2000[41] and, following the expiry of licences, TIFF-LZW have come into the focus of memory institutions as efficient compression formats. With regard to long-term archiving, the operators of repositories should carefully weigh the advantages and disadvantages of both formats (TIFF-LZW and JPEG2000). Compression formats are generally more susceptible to image loss and their use should be decided according to a risk-benefit assessment. When selecting a format, its prevalence and market penetration should also be taken into account. Regardless of the fact that large and influential libraries such as the Library of Congress and the British Library use JPEG2000, its prevalence so far is less than that of the uncompressed TIFF format.

The licensing situation is still not fully resolved, yet some parts of JPEG2000 have been declared free to use.

The various proprietary raw formats are also unsuitable for archiving master images, especially as they can often only be viewed with the corresponding raw software. The Adobe raw format DNG, which is not dependent on any particular platform or camera, has not entered widespread use and is therefore also unsuitable as an archive format.

JPEG and PNG are generally recommended for publication on the Internet due to their widespread popularity. However, other formats or techniques can also be employed if these are more practical for the presentation of, for example, very large image files. It is important to ensure that only formats and techniques are employed that can be used without special technology or software. For online presentation, it is especially important that the chosen format can be displayed natively by all common browsers.

For AV media, a distinction must also be made between archive and user formats. For audio data the Waveform Audio File-Format (WAV) in connection with Pulse Code Modulation (PCM)[42] has established itself as archiving format while MP3 (.mp3, MPEG-2 Audio Layer III) is the most popular user format. No clear recommendation can currently be made for video formats. However, care should be taken to use only licence-free and open formats. The problem with video files is their enormous size, which is why uncompressed storage is hardly feasible for larger projects.[43]

Digital 3D documentations of digitised or digitally reconstructed 3D objects are stored as 3D models. These models allow to accurately represent the form, texture and visual material characteristics of the original object, or to document and visualise objects that are no longer extant (e.g. destroyed).

Any data format that is taken into consideration should be as robust as possible against damage to the data medium, make efficient use of storage capacity, map the model with a logical structure, permit fast data processing and be in wide use.

OBJ (.obj)[44] has proven itself as open file format for 3D models. It is supported by many 3D graphics programs and is therefore suitable for the sharing of 3D models across programmes and platforms. Optical material properties (e.g. mirroring, transparency, highlights, etc.) are defined in a separate material file (.mtl), which may also contain

[41] ISO/IEC 15444-6:2013.

[42] Minimum quality: 44.1 kHz sampling rate and 16 bit sampling depth.

[43] The following formats are recommended for long-term archiving, although it should be noted that no satisfactory solution is yet available: MJPEG2000/MPEG-4 (ISO/IEC 14496-12:2015, ISO/IEC 15444-12:2015), DPX (standardised as SMPTE 268M-2003, v 2.0), MXF/AAF (standardised as SMPTE 377M and proposed as ISO standard).
FADGI is currently working on guidelines for archiving films:
http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html.

[44] http://www.fileformat.info/format/wavefrontobj/egff.htm and https://de.wikipedia.org/wiki/Wavefront_OBJ.

information on texturing. In addition, glTF[45] is used primarily in scientific contexts (also as a web format). The X3D (.x3d)[46] format, developed specifically for the visualisation of 3D models within WebGL technology, is suitable for web-based documentation and sharing. Due to its smaller scope of performance, however, it plays a comparatively subordinate role.

3.2.2 Material-specific parameters (incl. three-dimensional objects)

3.2.2.1 Text-based works

The term "text-based works" in this context refers to both printed works and rare documents such as manuscripts and archival materials.

Image digitisation should always be preferred for the digitisation of historical holdings. Even if a machine-readable full text is available, image digitisation and/or the presentation of the digital facsimile should not be omitted because a great deal of information can only be conveyed visually.

The legibility of the text is the deciding factor in the choice of resolution, and this depends on the size of the type or script. Hence, the required scan resolution depends less on the dimensions of the item and more on the legibility of the individual letters. For text documents where the smallest significant character is 1.0 mm or larger, a resolution of 400 ppi is recommended. A resolution of 300 ppi is only recommended if the minimum character size is 1.5 mm or larger.[47]

Folios / pages of a textual work are always digitised in their entirety with a thin surrounding edge to indicate that nothing has been cut off from the original.

3.2.2.2 Graphic representations

As described under (→ 3.2.1.1), the recommended minimum resolution for digitisation is 300 ppi, subject to the outcome of a test digitisation with a standardised test chart.

In the case of small items, however, a resolution of 300 ppi relative to the original format will often be insufficient to reproduce the characteristics of the original in consistently recognisable detail. If the artistic technique employed is not identifiable for example, a higher resolution should be chosen. Examples may include copperplate engravings, stamps, portrait medallions and miniature paintings.

The highest resolution that can be achieved depends on the technique and equipment used for digitisation. Here is a sample calculation, disregarding the lens quality: A stamp can be reproduced with a digital camera with a sensor in 24 x 36 mm format at a scale of 1:1.

Conversely, in the case of large items (DIN A0 or larger), the resolution can be reduced in relation to the original format if the objects are intended to be viewed from a greater distance. This applies to posters, for instance. In this case, the resolution can be reduced to as little as 150 ppi.

---

[45] https://www.khronos.org/gltf/.

[46] http://www.web3d.org/x3d/what-x3d and https://de.wikipedia.org/wiki/X3D.

[47] See Federal Agencies Digitization Guidelines Initiative (FADGI): Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files, August 2010: http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf, p. 59. https://www.digitizationguidelines.gov/.

However, if large-format items are very detailed, for example large topographical maps or copperplate engravings, a resolution of at least 300 ppi should be used. All image-relevant details must be recognisable and legible.

Due to the added user value, the full physical resolution of the capture technique should be used for graphic representations if economically viable.

Graphics are always stored in their entirety with a thin surrounding edge to indicate that nothing has been cut off from the original.

3.2.2.3 Photographs

A distinction must be made between transparent media (negatives or slides) and media intended for viewing (positive images, for example on paper). The latter are normally contact copies or enlargements of negatives on photographic paper.

Transparent media

A photographic negative or slide is the result of a photographic capture and therefore constitutes the original source. Negatives are used to make reproductions for disseminating the picture as a photographic positive or print.

Negatives are not suitable for viewing the subject. To make the subject usable, a positive derivative must be generated from the digital master. This serves as a digital print of the negative and reproduces the image content. However, the positive representation of the master is usually too dull and, in the case of colour negatives, not colour-true because of the colour mask. To clearly reproduce the subject in the digital print, considerable image correction is therefore required.

Well-prepared negatives yield images with a tonal range of up to 12 aperture stops, which must be reproduced during digitisation.

Unlike graphic representations, photographic images are generated technically. The quality of a photograph is thus dependent on a number of different technical factors: image format, lens, film type, granularity of the emulsion, development, exposure, focus and so on.

To take account of these aspects when choosing the most suitable scan resolution, the following alternative to the procedure described in section (→ 3.2.1.1) is a possibility.

To do justice to the various qualities within a mixed collection, a resolution of 80 lines/mm may be selected as a starting point to determine the scan size, which corresponds to the resolution of modern, fine-grained slide film. This ensures that the full detail of both historical and more recent photographs can be reproduced in digital form. In most instances, the resolution values of modern films are documented in manufacturers' data sheets. A scan resolution of 4000 ppi thus provides sufficient detail.[48] In terms of the film material, the theoretically achievable resolution would be constant for all output formats.

For small-format pictures, a scan resolution of 4000 ppi can be used. For larger image formats, the scan resolution can hence be reduced. A useful indicator is the maximum circle of confusion diameter of a dot, which is still perceived as sharp by the human eye for the particular format.

[48] To depict the film resolution digitally in lines/mm, at least two pixels are required per line. Thus, 80 lines/mm corresponds to 160 pixels/mm or 1600 pixels/cm. Multiplied by 2.54 (1 in = 2.54 cm), this gives a scan resolution of 4064 ppi, which can be rounded to 4000 ppi.

As a general rule for photographic sources, the format diagonal in mm (corresponding to the normal focal length) is multiplied by 1/1500. This gives the following maximum permissible circles of confusion for the various formats:

Small format: 0.03 mm; medium format: 0.05 mm; 9 x 12 cm: 0.1 mm; 18 x 24 cm: 0.2 mm

This in turn yields the following minimum target values, the physical resolution of the scanners or cameras should be fully utilised if possible:[49]

Medium format: 4000 ppi x 0.03/0.05 = 2400 ppi

9 x 12: 4000 ppi x 0.03/0.1 = 1200 ppi

18 x 24: 4000 ppi x 0.03/0.2 = 600 ppi

Photographs are always stored in their entirety with a thin surrounding edge to indicate that nothing has been cut off from the original.

Media intended for viewing

The photographic positive is an end product. Unlike a negative, it is not intended to serve as a starting point for reproduction or enlargement. The quality of the image therefore depends on the quality of the negative.

As described in section (→ 3.2.1.1), the recommended minimum resolution for digitisation is 300 ppi, subject to the outcome of a test digitisation with a standardised test chart.

Due to the significant added user value, the full resolution of the chosen capture technique should be used for transparent as well as viewing photographs if economically viable.

3.2.2.4 Microforms

For microforms (16 mm and 35 mm microfilm, 105 x 148 mm microfiche, positive and negative, B/W and colour), in terms of parameters the information given in the section on transparent media (→ 3.2.2.3) applies. However, with regard to resolution it should be remembered that these are reproductions of originals for backup or protective purposes. A reduction factor of between 1:7.5 and 1:96 may be used for filming. Wherever possible the resolution for digitisation should be based on the original and not the film.

The digitisation of microfilm is often used as a means of mass digitisation, which can be carried out at low cost. The recommended 300 ppi relative to the original size can only be achieved in very rare cases, even if the film's resolution is theoretically adequate. The limiting factor is the currently prevalent types of scanner, which can digitise whole films semi-automatically in a short space of time. For B/W microfilms and microfiche, it is possible to achieve resolutions of up to 600 ppi relative to the film. The choice of resolution must be based on the technical possibilities of the mass process. Reproduction techniques that allow high-resolution individual scans should only be used in exceptional cases.

If film microforms that conform to the standards of federal security microfilming (→ 2.3) are available and used as a basis for digitisation, it is necessary to check whether it makes sense to create and preserve a master. Microfilming itself is in this case a long-term storage medium and copy master. However, it must always be investigated whether

[49] See also: https://www.digitizationguidelines.gov/guidelines/digitize-technical.html. General FADGI Guidelines under http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf (Federal Agencies Digitization Guidelines Initiative (FADGI): Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files, September 2016).

it makes more sense to create digital copies (master and user versions) of the original than to generate versions for users from the microfilm.

3.2.2.5 Three-dimensional objects

The digital capture and sharing of 3D objects in the cultural heritage sector has considerable significance for a number of disciplines and enables new methodological analyses and insights. Digital 3D documentation captures all research-relevant objects, for example through the 3D digitisation of cultural heritage objects (such as museum collections, architecture and its features, etc.).

In the past, this was done mainly by taking photographic images from different angles whereas today, a proven selection of 3D capturing and reconstruction methods as well as technologies supporting these processes are available.

In contrast to photography, 3D documentation captures, where possible, the entire geometry of an object, yet always its surface texture and, where possible, its visual and physical material properties, integrates and consolidates them in the digital 3D model.

3D digitisation provides a more comprehensive recording compared to a photograph, among other things, by capturing shape and material properties. This allows the surface-light interaction of the object to be recorded and reproduced faithfully via manual post-processing (with a suitable renderer). The 3D model can be visualised and simulated with the appropriate software from any perspective in the lighting situation of when the photo was taken, as well as in new lighting situations and environments.

There are many tools and game engines for the interactive web presentation of 3D models, which are available as 3D web viewers through commercial providers (e.g. Sketchfab) or open source (e.g. Babylon.js). In addition, freely available solutions have been created specifically for scientific purposes. These include the so-called 3D Heritage Online Presenter (3DHOP) as well as the tool Kompakkt (which is currently being further developed as part of NFDI funding, in this course enabling the display of point clouds). They offer, to some extent, specific solutions as well as a repository. Moreover, the DFG 3D Viewer, which is currently being built, will provide another solution.

For online representation, the 3D model should be available in a suitable form (precise specifications regarding polygon, amount or file size cannot be generalised. Formats result from the tool that has been used, usually .gITF or .obj). When it comes to digital 3D documentation, this document is concerned with retrodigitisation (transfer of a physically existing object to a digital copy and 3D models derived from it). On this basis, reconstructions of a physically non-existent object in the form of a native digital 3D dataset become possible, which will not be further discussed here.

Digital 3D models are characterised by a number of advantages, given as examples below:

- Digital 3D models of cultural objects can be made easily available and can be accessed by many researchers at once (even in different qualities)
- Missing (e.g. destroyed) parts can be added in reconstruction
- Different hypothetical versions and/or variants of a 3D object can be simulated
- Digital 3D models can be used by museums for exhibition planning, documentation, acquisition planning etc.
- Digital 3D models can be used for virtual presentation and exhibition to the public (in combination with new presentation techniques such as hybrid exhibits) in order to transfer knowledge and increase attractiveness to visitors

- Digital 3D models are a suitable reference for the restoration of damaged originals / generation of physical replicas on the basis of the digital 3D model
- Digital 3D models may be a substitute for loans (avoidance of damage, insurance costs, legal uncertainty relating to ownership)

Data formats for 3D models depend on the mathematical form of representation, which is also influenced by the recording or creation process. This specifically refers to the difference between meshes and point clouds. Furthermore, models that are archived for the purpose of long-term preservation are to be distinguished from web representations. Proprietary file formats should generally be avoided in order to ensure software-independent use of the models. The resolution of the textures, the number of polygons or points etc. results from the purpose of use and the intended representation.

The 3D repositories run by Mainz University of Applied Sciences and Heidelberg University Library provide free infrastructures for publishing 3D models. There is also Sketchfab, a non-free platform (costs, rights) for publishing, sharing, buying and selling 3D, VR and AR content. There is currently no low-threshold documentation and exchange format. In this regard, results are to be expected in the near future in the form of an application profile for encoding and descriptive metadata stored in MODS.

There are various optical digitisation methods for the retrodigitisation of 3D objects:

- Laser scanning methods are divided into run length measurement and triangulation. In the first case, the time it takes a laser pulse to travel to the object and back again is measured in order to determine the distance to this point. In the second case, the three-dimensional surface of an object is extrapolated from the distortion of a projected laser line as captured by a camera.
- Strip light methods use the projection of defined patterns (mostly parallel lines) on object surfaces and likewise extrapolate 3D surface geometry from distortion. To increase resolution, these patterns can be phase-shifted across surfaces.
- Photogrammetric methods involve taking pictures of the object to be digitised from different angles and scanning by pairs of features for at least two overlapping image recordings, so that the depth of the associated image point is obtained by triangulation. Characteristic features may be, for example, differences in intensity or contrast.

There are different technologies to choose from, depending on the type of object and its properties (size, material, condition, etc.). Currently, the project Colour and Space in Cultural Heritage[50] (COSCH) is trying to develop an online guide for selecting an adequate methodology and technology depending on the task. However, making early contact with leading technology developers and users in the area of mass digitisation continues to be important.

The principle of time-of-flight cameras is similar to run length measurement in laser scanners, with the difference that an entire scene is illuminated and photographed at once, and a depth value is calculated for each pixel of the camera sensor.

For objects above a certain size, for instance large statues or buildings, a point-by-point measuring technique such as laser scanning or surface coding is no longer efficient, and photogrammetric techniques demonstrate their strengths.

[50] http://cosch.info/.

In retrodigitisation, there is a general relationship between object size and resolution or, more precisely, between the measurement range of a sensor and accuracy. The larger the object or the greater the distance, the less precision with which details are recorded.

Geometry and texture are normally recorded in the same step as geometry acquisition, because there is a direct correlation between the measurement points and the points of the imaging system. Many acquisition methods use a camera sensor as an integral component of image capture to extract depth information. This sensor usually has the best viewpoint for each captured section of the object, so the same sensor position can be used for texture capture.

To capture the texture, sufficient, diffuse and even lighting is required. The right white balance is also important, as the mostly artificial diffuse lighting on the object surface must be compensated for. Colour calibration is essential here as texture has a major influence on realistic appearance.

In addition to geometry and texture, optical material properties, such as gloss and reflection, can also be captured. This is typically done by recording combinations of light incidence and perspective onto an object in order to subsequently store the textures relevant for each technically defined surface point (visibility test). As a result, when the virtual object is later viewed interactively, the physically correct textures can be applied as well as possible under different lighting conditions, thus coming as close as possible to the original material appearance and light-surface interaction. In some cases, an object cannot be captured in its entirety, either because of its size, or because only remains of material are extant. In such cases, selective capture of material properties and subsequent extrapolation to regions of the same or similar material may be appropriate.

An important contribution to integration and convergence between digitised and native digital data is made by structured data geared to relevant metadata schemas (e.g. LIDO and CARARE 2.0[51]), controlled vocabularies (e.g. Getty Vocabularies[52]) and authority files (e.g. GND) as well as reference ontologies (especially CIDOC CRM[53]). The extension of CIDOC CRM for better capturing of retrodigitisation was implemented in CIDOC CRM_{dig}[54]. Application ontologies for knowledge representation of computer-assisted reconstruction are currently being introduced to and discussed by the community of experts.[55]

3.3 Metadata

The generation of metadata to enable the location of objects and the contextualised presentation of their digital representations is a central aspect of digitisation. The DFG assumes that the analogue objects selected for digitisation have already been indexed in recognised digital indexing systems and/or will be indexed in detail as part of the digitisation process. Metadata produced as part of the digitisation project must always be made available in software-neutral and standard-compliant form, usually XML encoding. This task should be integrated into the project workflow in such a way that, even if the project is terminated early for any reason, a complete set of metadata will be available in a software-neutral format.

If the objects selected as part of a funded digitisation project are factually suitable for inclusion in a material- or subject-specific portal, the project proposal should point out the approach that applies: either outlining what

[51] https://pro.carare.eu/en/introduction-carare-aggregation-services/carare-metadata-schema/.
[52] https://www.getty.edu/research/tools/vocabularies/.
[53] https://cidoc-crm.org/.
[54] https://cidoc-crm.org/crmdig/.
[55] http://www.digitale-rekonstruktion.info/.

measures will be taken in the project to facilitate a data integration to this portal during and after the project, or explaining why a data integration is not necessary or desirable for reasons of content or time and expense.

In general, a distinction is made between descriptive metadata (bibliographic description, archival indexing, description of rare (frequently non-textual) objects), structural metadata (text or document structure), administrative metadata (e.g. rights management) and technical metadata (e.g. file types or technical image metadata in EXIF format). The following considerations apply only to descriptive and structural metadata.

The link between an object and metadata on the one hand and between an object and its digital representations on the other hand must always be assured at metadata level. In addition, metadata may also be embedded in the headers of master files. However, since these are displayed differently by different software applications and may in the worst case even be corrupted, this is only a supplementary option.

Metadata formats reference the digital images in various ways. The container format METS allows associating descriptive metadata via mappings in standard formats (e.g. MODS, TEI), as well as structural metadata including references, to digital derivatives. This is especially suitable for fully digitised text-based works that are indexed with both descriptive and structural metadata. However, METS is by no means limited to digitised images and can reference any digital resource. Other formats like LIDO include, in addition to semantic elements for object description, specific elements for referencing digital resources for the object (besides digitised images also audio or video data). Any number of resources can be linked with an object and still have their own descriptive elements. The common EAD profiles allow an object to be linked with an unlimited number of digital resources, including referencing of METS containers, as well as other formats for referencing digital copies or even full texts. It is important that digital resources are referenced with globally unique persistent addresses, generally URLs, within the metadata (→ 5).

For library and archive material, digital copies and metadata must be recorded either by cataloguing the electronic version or by giving the PURL of the image files or a persistent link in the catalogue (OPAC, network system, information system). Applicants from university institutions are expected to coordinate cataloguing with their local institutions or actually have it done by them.

Digital copies should be listed in relevant material-specific portals (e.g. Handschriftenportal[56], Archivportal-D or Kalliope[57] for remains). Materials that cannot be added to library or archival indexing systems should be presented in suitable subject-specific or multi-subject online applications. Digital copies and metadata should be listed in the DDB. All projects are expected to contribute data to the DDB and from there to Europeana. Data should be delivered to these portals in accordance with standard formats, if possible via OAI (→ 3.3.4).

It is up to the funding recipient to ensure that the digitisation files (image and text) produced by the project can be uniquely identified as well as searched and retrieved separately from other units recorded in the same system.

3.3.1 Indexing, descriptive metadata

The indexing of an object with descriptive metadata fulfils different functions depending on the material. A distinction is made between descriptive metadata for display in the DFG Viewer and those for describing the

[56] https://handschriftenportal.de/.

[57] http://kalliope.staatsbibliothek-berlin.de/de/index.html.

content of the object in the portal. The descriptive metadata is what makes it possible in the first place to locate the content by a search. The object is classified and placed in its historical context, and, if applicable, its materiality is described. Ideally, descriptive metadata should provide links for different enquiries and disciplines. Funding is not available for digitisation projects that do not include metadata classification in line with established community standards.

For minimal indexing, the DFG Basic Data Record[58] can be used. It contains all necessary mandatory and recommended data fields for describing metadata, analogue source objects and their reproductions. If a digitisation project is intended to include project-specific indexing and/or information provision beyond descriptive metadata (which will usually be the case for rare objects at least), the most important enquiries relating to the material are to be anticipated. Additionally, these should be formulated in a core field catalogue using the relevant mandatory fields and standards (and, if applicable, recommended or optional data fields). The systematic preparation of metadata based on a core field catalogue is essential to the optimum provision of digitised holdings (→ 6.2).

To ensure the optimally distributed and long-term usability of metadata, indexing should be based on relevant standards (e.g. RDA[59], ISAD(G)[60], RNAB[61] and reference models (CIDOC-CRM[62], possibly also IFLA LRM[63]/FRBRoo[64]) and linked with published authority data wherever possible. To record personal biographical information and for geographic indexing we expressly recommend the Integrated Authority File (GND) offered by the German National Library.

Other controlled vocabularies such as Iconclass[65] for image classification, the Getty Art and Architecture Thesaurus (AAT) or specialised vocabularies such as GeoNames, Getty Union List of Artist Names (ULAN) etc., should enable national and international links.[66]

Metadata must be made available for further use in accordance with material-specific standards: METS/MODS for manuscripts and printed text-based works (→ Appendix A), METS/TEI as additional option for manuscripts (→ Appendix B), a common EAD profile[67] in connection with METS/MODS mapping for archive material (→ Appendix A), LIDO for (usually rare) pictorial and three-dimensional objects (→ Appendix C). A METS/MODS profile is currently being developed for 3D models (which can be used in conjunction with LIDO), multimedia AV media and music. The metadata must be valid against the relevant XML schema and must also be checked for semantic correctness.

[58] https://www.dfg.de/formulare/12_155/.

[59] https://wiki.dnb.de/display/RDAINFO/RDA-Info.

[60] http://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_DE.pdf.

[61] https://d-nb.info/1186104252/34.

[62] CIDOC Conceptual Reference Model: http://cidoc-crm.org/.

[63] IFLA Library Reference Model: https://repository.ifla.org/handle/123456789/40.

[64] IFLA Functional Requirements for Bibliographic Records - object-oriented: https://repository.ifla.org/handle/123456789/659.

[65] https://www.iconclass.org/.

[66] https://www.getty.edu/research/tools/vocabularies/aat/; https://www.geonames.org/; https://www.getty.edu/research/tools/vocabularies/ulan/.

[67] https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=19010180. The DDB plans to move closer towards the international standard EAD (3.0) over the next few years.

3.3.2 Structural metadata for digital facsimiles

It is worth considering the use of structural metadata to index digitised text documents, i.e. encoding the structural elements of a document such as dedications, prefaces, chapters or illustrations (→ 3.4.3). In some cases the production of structural metadata is of somewhat lesser importance, whereas for extensive compendia, dictionaries or similar, it is essential to create an artificial table of contents to enable the user to navigate within the digital document. For some enquiries, structural metadata are also of interest for search purposes. Therefore the decision whether to generate structural metadata is always material- and project-specific.

If structural metadata is used, it is recommended to consult the structural data set available on the DFG Viewer website, a list of designations supported by the DFG Viewer.[68] If additional designations are needed, standardised terms for a given digitisation project should be agreed on. This typically specialised vocabulary should be published on the project's website and, if appropriate, the DFG Viewer website to allow others to reuse it.

When assigning structural metadata, the question arises as to whether document indexing should follow the digital facsimile, the physical page sequence, or the work's text and/or chapter structure.

If a transcription or edition will accompany the digital facsimile of an old print or manuscript, the recommended encoding standard is TEI. For a page description with some qualifying features (e.g. illustrations or annotations), the standard METS, according to the application profile for using METS/MODS in the DDB[69], is recommended. There are good arguments for both the page-oriented and the document-oriented model. Usually it is even possible to merge both approaches. A structure that follows the logic of the source text tends to be more powerful in terms of later retrieval options and digital representation. However, this advantage comes with the price of greater technical requirements for processing and displaying the digital representations.

It should be pointed out that even encoding based on the physical page sequence, which tends to be more common in libraries, does not rule out the use of TEI. For the synchronisation of full text / transcription and digital copy down to page level, structural data should always be recorded in METS, regardless of whether additional structuring in TEI is carried out or not.

The standards currently recommended for old prints are METS or TEI. If TEI is used for structural data, the project must also convert it to METS in order to use the DFG Viewer.

3.3.3 Description of collections and holdings

Descriptions of collections and holdings contextualise and locate the individual object, thereby enabling the user to get an overview of the overall holdings of an institution. The basis for a digitisation project can also be virtual holdings and collections (e.g. subject-specific, material-specific).

Digitisation projects are expected to present at least the nature and scope of the selection of materials and/or objects on a website, preferably with an English-language version. A standardised description is desirable to facilitate the future merging of this information in national or international portals that enable the respective retrieval. This description may be based on the same metadata standard in which the object descriptions are made available: MODS or TEI headers, the common EAD profiles and LIDO all offer the necessary features. The RDF-based

[68] https://dfg-viewer.de/en/structural-data-set.

[69] https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=19006651.

Data Catalog Vocabulary (DCAT) is also becoming increasingly relevant.[70] In addition, a unique identification and description according to the ISO 27730 standard, Information and Documentation – International Standard Collection Identifier (ISCI), which is based on an institution's ISIL,[71] should be considered.

3.3.4 Exchange and dissemination of metadata

For the development of comprehensive reference systems, it is crucial to create a global standard for exchanging metadata of digital copies (→ 6). However, standards can be developed and established only within the respective community. The same resources may well be relevant for entirely different research questions. Accordingly, digital resources require diverging sets of metadata. A generalised procedure for exchanging metadata must therefore be flexible enough to handle different metadata formats and community-based specifications. This can be achieved well with the OAI protocol. With regard to old prints and manuscripts, OAI is especially useful as a technical exchange protocol. OAI requires that Dublin Core data be provided as a minimum. While Dublin Core data are insufficient for content-oriented descriptions of old prints and manuscripts[72] as well as non-textual resources[73], they are useful as additional information. The OAI standard explicitly enables the parallel support of additional metadata formats, so that OAI generally can be combined with all XML-based metadata formats (METS/MODS with sector-specific mappings, METS/TEI-HSS, MARCXML, MABxml, common EAD profiles, TEI P5, LIDO, etc.).

The DFG requires the provision of metadata via OAI as interface of the DFG Viewer. In addition to the Dublin Core metadata required by OAI, material-specific metadata must be provided in METS/MODS for old prints, METS/MODS for archive material with a mapping addressing the DFG Viewer, METS/TEI for medieval manuscripts, and LIDO for pictorial and three-dimensional materials (→ 7 and Appendices). At this point, however, it makes sense to point out the IIIF Presentation API[74] as equivalent possibility for making data available. Provision of metadata can also be realised via a portal (e.g. DDB, Archivportal-D). Any deviation from this requirement must be explicitly justified in the proposal.

3.4 Full text generation

A versatile scientific reusability of digitised texts which permits automated searches, quantitative evaluations with text or data mining, semantic analysis, pattern recognition in non-textual materials, enrichments, contextualisations and further processing – in both physical and virtual research environments – depends on the easy and unhindered usability of the data, appropriate legal permissions and the availability of the full digitised text. Whenever feasible and reasonable, DFG funding for text-based works should be used to generate machine-readable full texts. All proposals for digitising text-based material are therefore expected to address the possibilities of full-text provision. The OCR-D Coordination Project,[75] which provides techniques of OCR-supported full-text capture, specifications and guidelines while developing sample workflows, should be considered when planning full text recognition via OCR. The results of this project will be incorporated into the next version of the

[70] https://www.w3.org/TR/vocab-dcat-3/. See also Data Catalog Vocabulary (DCAT) - Version 2. W3C Candidate Recommendation 04 February 2020 https://www.w3.org/TR/vocab-dcat-2/.

[71] https://sigel.staatsbibliothek-berlin.de/vergabe/isil.

[72] See Hillmann, Diane I.: Choices: MARC or Dublin Core? In: Anne R. Kenny/Oya Rieger (eds.): Moving Theory into Practice. Digital Imaging for Libraries and Archives. Mountain View: Research Library Group 2000, p. 89f.

[73] The impermissibly simplified description of museum objects in Dublin Core and their representation in online environments was the main reason why the museum community developed the harvesting format LIDO.

[74] https://iiif.io/api/presentation/3.0/.

[75] OCR-D: Coordination Project for the Development of Methods for Optical Character Recognition (https://www.ocr-d.de/).

Practical Guidelines. Please note: For printed works dating from 1850 or later, the full text must be generated in addition to image digitisation.

Full text can be generated through manual or automatic processes: either via OCR/HTR or transcription. The procedures differ in their approach as well as in other areas such as format and quality of the result. The question of which method to choose needs to be decided with regard to the respective requirements and depending on the age and condition of the master copy.

3.4.1 Text capture

Text accuracy

Whether full text is being generated via OCR/HTR or manual transcription, it is important to decide what quality is needed for what purpose and what costs are appropriate. Text quality requirements will vary from one project to another. What standard is considered adequate and what costs are acceptable for what quality level must be carefully justified on a case-by-case basis depending on the material and scholarly requirements. The quality of OCR-generated texts is usually given as a percentage. There is no consensus as to the measurement criteria and processes to be applied. Accuracy may refer to the correctness of individual letters or of whole words. In the first case, 99% means that 1 in 100 letters is wrong, and in the second case 1 in 100 words. Whether incorrect layout information (marginal notes correctly recognised but placed in the wrong position) or missing syllabification are considered errors depends on the project and its criteria.

To create some uniformity on which to base the evaluation of accuracy, applicants are requested to state the letter accuracy, i.e. incorrect syllabification and layout errors should be ignored. Applicants should therefore check how many characters in the source text were correctly recognised, including punctuation characters. Ideally, measurements should be made on the basis of reliable reference texts, but as these are not always available, random sampling may need to be used.

For now, a statistical method is recommended to check the accuracy of transcribed or OCR-generated texts. Future findings, e.g. based on the results of the OCR-D project, might lead to a different recommendation. Statistical methods should be used to assess whether the recognition rate claimed by a service provider can be relied upon. The tests should be based on a random sample. The probability of error should be kept as low as possible while keeping the size of the random sample manageable. The recommended statistical method is the so-called Bernoulli trial. As the calculation is relatively complicated, we recommend using the following specification. The recommended basis is a random sample of 500 characters. To select the position of the characters we recommend the use of a random generator (1st character: page 15, line 24, character 7. 2nd character: page 73, line 3, character 32 etc.). To check the accuracy, the following table applies:

Claimed recognition rate	Minimum number of correctly recognised characters (random sample size = 500).
95 %.	485.
96 %.	489.
97 %.	493.
98 %.	496.
99 %.	499.
Over 99 %.	500.

The left column shows the claimed recognition rate. The right column shows the minimum number of correctly identified characters that must occur in the random sample to verify that a claimed recognition rate is correct. Hence, if a service provider claims a text has an accuracy of 96%, at least 489 characters must be correctly recognised in the random sample of 500 characters with a 2.5% probability of error in order for the service provider's claim to be accepted. An accuracy of less than 95% should preferably not be agreed upon.

If the recognition rate is claimed to be more than 99%, the size of the random sample would have to be increased. The two tables specify the minimum number of correctly recognised characters in relation to the size of the random sample for texts with claimed recognition rates of 99.5% and 99.7%:

Claimed accuracy: 99,5 %

Random sample size	Minimum number of correctly recognised characters
500	500
1000	999
2000	1996
5000	4985
10000	9960

Claimed accuracy: 99,7 %

Random sample size	Minimum number of correctly recognised characters
500	500
1000	1000
2000	1998
5000	4995
10000	9990

It goes without saying that there are pragmatic limits to sampling. The question as to what level of accuracy is good, sufficient or insufficient depends on the use of the text and the concrete project requirements.

OCR and HTR

OCR/HTR technology has made considerable progress in recent years. Significant improvements have been achieved on blackletter fonts, both from the hand press and the machine press era as well as with regard to manuscripts. However, considering the dynamic development and the results expected particularly from the OCR-D project, the Practical Guidelines cannot make any conclusive recommendations concerning OCR/HTR software and its usability for now.[76]

[76] For computational recognition of historical printed works, different font types and manuscripts, for instance, the free software OCR4all was developed, see http://www.ocr4all.org/. There is now a connection with the approaches developed in OCR-D.

The OCR/HTR process itself consists of the following stages that are based on each other: Preprocessing to prepare the image (cropping, despeckling, deskewing, binarisation), Optical Layout Recognition[77] (OLR = segmentation or layout analysis = identification of image and text parts, structural analysis), Optical Character Recognition (OCR = actual text recognition) or Handwritten Text Recognition (HTR) and postprocessing (text correction). The quality of the original image has a decisive impact on the binarisation process, which in turn is the basis for quality of character recognition; for this reason, only digital images of sufficient quality should be used for OCR processing. Problems may arise as a result of intrinsic phenomena such as dirty marks, shadows of the reverse side, manual underlining and annotations, which could have an adverse effect on the OCR/HTR process. Based on binarisation, the software identifies the text areas in a digital image and separates the actual text from illustrations or other pictorial elements (= segmentation). Complex layouts such as marginal notes in manuscripts, older printed works or newspapers as well as other multi-column documents may significantly interfere with the segmentation process. Binarisation and segmentation are the basis for the actual OCR/HTR process of text recognition, which may be supported by a subsequent process of text improvement (automatic use of dictionaries or manual correction).

The ALTO (Analyzed Layout and Text Object) XML standard, which is maintained by the Library of Congress, is recommended to facilitate the reuse of data from OCR/HTR.[78]

Manual capture/double keying

There are two methods used to transcribe texts: the single-key and the double-key method. In the latter, a text is transcribed twice, then the two versions are compared automatically and any discrepancies are filtered out. This allows for transcription accuracies of up to 99.97%, i.e. virtually error-free texts.

When choosing this type of transcription, one should not be misled by service providers claiming ostensibly high accuracy rates; results with less than 99.5% accuracy are inadequate for manual transcription.

If transcription is to be outsourced to a service provider, the contract must specify the contracted accuracy rate. Adherence to the contracted rate should be verified by random checks of the digitised text (see above).

Manual transcription has the advantage of high accuracy but is costly. Advantages and disadvantages compared to OCR/HTR capture must be weighed. The manual transcription is often done outside of Germany; however, a contracted digitisation provider should have a representative in Germany since close cooperation and discussion on transcription details are usually needed.

As a first step, a digitisation project must determine which properties of the master copy should be captured by a structural markup. Only features that are graphically distinct can be marked up. Simple structures can be recognised automatically by the service provider; further details must be marked in the images before the materials are handed over to the contractor. This requires a certain capacity of personnel, which must be taken into account when calculating the project budget.

Because most service providers invoice based on the number of characters including markups, it is advisable to use a markup language with few characters[79] for this purpose.

[77] The term "optical layout recognition" was originally coined by Fraunhofer IAIS for the marketing of their OCR software myDec and has since become somewhat independent. In scientific discourse, on the other hand, "(document) layout analysis" or "segmentation" is more common.

[78] http://www.loc.gov/standards/alto/.

[79] e.g. TEI Tite: https://tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

3.4.2 Character encoding

All common operating systems support Unicode. Unicode is also the character encoding format for XML, which forms the basis of the most important structural data markup systems.

It is therefore recommended to save the texts in Unicode. Preference should be given to UTF-8 (without Byte Order Mark BOM), which is more economical with European languages. Characters which are not part of the Unicode standard can be depicted by using the private plane area of Unicode[80] and represented by appropriate graphics or fonts. Before defining own characters in the private use plane of Unicode, it needs to be checked whether characters are included in the community based MUFI standard[81]. In all cases, options for standardisation should be explored.

The question of whether to encode the long and short s in blackletter fonts, ligatures in blackletter fonts (ch, tz etc.) and diphthongs (æ etc.) depends on subject-specific requirements or editorial decisions. These are beyond the scope of these recommendations, but should be borne in mind when generating full text and should ideally be documented.

3.4.3 Markup of full texts

Unless there are cogent reasons not to, full texts of prints and manuscripts must be encoded and marked up following the model of the TEI. As a transparent XML format, TEI is also prospectively the best choice for long-term archiving as long as it is thoroughly documented. PDF/A should be avoided, in spite of existing ISO standards (19005-1:2005 and 19005-2:2011), since its use, especially in the digitally working humanities and cultural studies, is limited by the lack of structural markup.

However, as a derivative format, PDF – and increasingly ePub for mobile devices – is suitable for dynamically generated reading versions or print-ready text, and should be offered as an extra option by digital libraries because of its widespread popularity (see for example the material available at archive.org).

When encoding XML structures in TEI documents, it must first be decided how, and to what extent, text-type specific divisions such as volume, article, chapter or section etc. should be taken into account. The same applies to other possible structural features such as tables of contents, indexes, line breaks, column breaks, page breaks, headers / footers / running titles, page numbers, images or image-like elements, captions, marginal notes, but also changes of font, e.g. a change from font type Fraktur to Antiqua (e.g. for foreign quotations), changes of font size, changes of font style (normal, italic, bold etc.), and others, formulas, e.g. mathematical (MathML) or chemical (CML) formulas, embedded code, continuation marks (catchwords) at the bottom of a page (referring to the following page), and more.

The choice of markup is generally dependent on the particular project. To ensure that full texts, marked up in this way, are exchangeable and reusable, the XML elements and attributes used should be documented in the TEI header. Please note the efforts of the TEI community to develop a strictly formulated exchange format: TEI-Simple.[82]

---

[80] https://en.wikipedia.org/wiki/Unicode.

[81] https://mufi.info/m.php?p=mufi. Before defining your own characters in the private use plane of Unicode, you should also check whether a standardised encoding already exists within the framework of OCR-D encoding (https://ocr-d.de/de/gt-guidelines/trans/ocr_d_koordinationsgremium_codierung.html).

[82] https://github.com/TEIC/TEI-Simple.

3.4.4 Layout

In some cases, when presenting a full text, it is important to preserve the layout of a document for the long term. The Practical Guidelines recommend using a suitable formatting language (e.g. XSLT, CSS), which largely ensures independence from special software. If valid reasons prohibit archiving the format with XML techniques, layout information about text documents may also be archived in PDF according to ISO standard 19005-1. However, as explained in 3.4.2, PDF files are no substitute for providing the marked-up full text in XML, as they cannot be evaluated in a structured way. Besides TEI, richer XML formats like PAGE XML and the mentioned standard ALTO XML (Analyzed Layout and Text Object, ALTO v4) etc. have meanwhile established themselves for layout information and are also recommended.[83] For presentation purposes, these can be transformed into HTML/CSS whereas, however, source data should be stored in the first-mentioned formats.

A publication in XML + formatting language allows dynamic views to be generated according to the user's intended purpose. This should be taken into account in the presentation and as wide a spectrum as possible should be offered. Typical output formats of XML files include HTML/XHTML, PDF, ePub and plain text.

3.5 Long-term availability

Long-term preservation and archiving of digital content is solved technically and organisationally, is well documented by models and auditable based on standards.[84] A distinction is made between archiving / preserving genuine digital information and (retro-)digitised copies of existing analogue objects. Different criteria can be applied regarding these types of archiving.

For long-term preservation, files are stored in stable, migratable formats in a technically and organisationally secure storage system (bitstream preservation).

Digital data is archived within a similar storage system but with more extensive technical and organisational measures that cover not only the physical retention of the data, but also strategies for enabling use (access) of the data, also in the context of existing and future information systems. The storage systems must be designed for redundancy and be audit-proof.

For archiving, both technical information and information about the change history of an object are particularly important. Especially with regard to the change history of an object, PREMIS (Preservation Metadata: Implementation Strategies)[85] is the data model of choice.

Whether long-term preservation or archiving is chosen for generated digital content will depend on the strategy of the institution, which must be outlined in the proposal.[86] Key criteria for successful long-term preservation and archiving of digital documents are as follows:

1) Creating the necessary organisational and financial framework,
2) Creating the necessary technical environment and choosing suitable techniques and strategies.

---

[83] See above section (→1.4), notes 19 and 20.

[84] For a good overview: https://www.langzeitarchivierung.de/Webs/nestor/EN/Publikationen/publikationen_node.html.

[85] http://www.loc.gov/standards/premis/understanding_premis_german.pdf.

[86] For example, an institution may specify in its strategy that born-digital objects should be archived and digital images of analogue objects should be stored in long-term storage systems. It may also be advisable to archive the digital copies of the original if the originals are fragile, for example.

The Open Archival Information System (OAIS) should be used as a reference model for the archiving of digital data.[87] The "Criteria for Trusted Digital Repositories" are essential. They define organisational framework, legal basis, quality management and authenticity for a trusted repository.[88] Long-term preservation is implemented in digital magazines or reproduction administrations based on the OAIS model.

The long-term availability of the results of digitisation projects depends firstly on the choice of data and metadata formats. These have been outlined in previous sections. Secondly, it must be ensured that the digital data remain physically available. It should be noted that in DFG-funded digitisation projects, the costs of project-specific data preservation are expected to be borne by the institution for the duration of the project. DFG funding is not available for these costs.

Long-term preservation and archiving is an integral part of any digitisation project. The costs and efforts it entails should not be underestimated. The long-term costs of storage, which, depending on the project, may amount to several terabytes, and the long-term effort involved in physical retention must both be taken into account.

It should be noted that the DFG views digitisation projects as endeavours undertaken by the entire institution: It is assumed that the department carrying out the project will be supported by the in-house IT infrastructure. We also encourage smaller institutions to take advantage of the expertise and services of larger institutions.

Proposals must contain convincing statements regarding institutional long-term preservation and archiving.

[87] The OAIS reference model has been approved as an ISO 14721 standard: http://public.ccsds.org/publications/archive/650x0m2.pdf. German translation at: http://nbn-resolving.de/urn:nbn:de:0008-2013082706.

[88] See Schoger, Astrid/Susanne Dobratz/Reinhard Altenhöner: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Frankfurt am Main, 2008. See: http://nbn-resolving.de/urn:nbn:de:0008-2008021802. See also: DIN 31644:2012-04.

4. Organisational Issues – In-House vs. Outsourced Digitisation

Digitisation may be done in-house or outsourced. In the first case, in addition to personnel costs, equipment costs can be funded if they are project-specific and not part of the core infrastructure to be provided by the institution.

The decision whether to undertake digitisation as an in-house project or to outsource it is always specific to the project, taking into account all practical and funding considerations, and exclusively the applicant's responsibility. Using contractors for direct digitisation is, above all, a matter of trust. Unlike with film digitisation, where the originals are not at risk, the service providers hired to handle rare material or historic library holdings should have the appropriate experience.

Outsourcing is possible even if the materials to be digitised may not be taken off the premises; on larger projects, service providers will often work on site, bringing their own personnel and equipment.

The following issues should be considered when drafting agreements with suppliers:

- Job parameters must be exactly specified, in particular the requirements and format standards for deliverable raw data. These must be formulated in a project-specific specification in accordance with the "Practical Guidelines on Digitisation". Contracted suppliers should be able to demonstrate certified quality assurance procedures. The outsourcer is required to perform careful quality control on deliverables before settling invoices in full.
- The DFG expects that an appropriate percentage of the invoice amount be withheld for security purposes and not paid out to the business providing the service until a quality check has been performed. In addition, the business should be required to provide a written undertaking that it will, without delay and free of charge, render substitute performance or rectify defects should this become necessary due to its non-compliance with the specification or other justified quality complaints.

5. Citing Digital Resources, Persistent Addressing

When digitisation was in its infancy, the issue of citability of digital resources was frequently underestimated. However, it is citability that makes Internet-based digitised sources viable for academic use. In contrast to previous secondary formats, like microfilm or paper printouts, an Internet resource is not just a copy of the original, which can be treated and hence quoted like the original, but rather an independent object in a dynamic integral research environment. When a digital resource is online, it needs a unique address, so that other documents or databases can link to it. In addition to the customary analogue citation format, which can and should still be given via the navigation software, this requires the specification and online documentation of addressing techniques according to the granularities required in a research project (book, page, collection, file, object, object part, etc.)

A positive effect of the net-based citation format is that referencing becomes unequivocal – something that usually cannot be said about old prints, because of the many mistakes they contain, or documents like incunabula, which lack pagination or foliation. Therefore the "content-based" citation format (e.g. p.8, a4, 213r etc.) should be joined by a "formal" citation based on the image sequence.

This also allows for the unequivocal citation of digital copies that are not part of the work proper (cover, endpaper, additional digitised watermarks, partial reproductions of illustrations etc.). Other mechanisms apply to full text, for which no specific recommendations can be made as yet (options for unequivocal referencing include, for example, XPointer, ID issuing and similar techniques).

For printed works and manuscripts, the accessibility and citability by means of URI or IRI of the work as a whole as well as of its individual physical pages must be guaranteed. Institutions should implement suitable mechanisms (PURL, URN, DOI, Handle etc.) to ensure the persistence and linkability of a resource, thus reliably providing sources for scientific research. Important to point out is also the benefit of unambiguous citability by URI of resources in terms of the further development of the Semantic Web and Linked Open Data (LOD).

It is strongly recommended to generate URNs at least at the work level via the German National Library.[89]

6. Provision of Project Results to the Public

6.1 Rights, licensing and open access

The DFG funds the digitisation of scientific material in order to make this material available to researchers in Germany and around the world. All projects should be designed to make results available to the research community promptly and for the long term. In virtually all cases, this will entail the provision of digital copies on the Internet. If, for legal reasons, metadata and/or digital images cannot be freely made available via open access soon after their creation, it must be ensured that this will occur promptly after any legal restrictions cease to apply.[90]

The DFG is a cosignatory to the Berlin Declaration on Open Access.[91] In the spirit of this declaration, the results of DFG-funded digitisation projects should generally be accessible free of charge to researchers around the world. If a project will deviate from this requirement, specific reasons must be stated in the proposal. Restrictions to open access to the generated data may be justified for reasons such as privacy, copyright or archival law, but may not affect more than 10% of the total material to be digitised.[92]

6.1.1 Procedure prior to publication

Prior to digitisation, the rights to the material to be digitised must be determined and, if necessary, obtained for digitisation and provision (→ 1.8; 2.2). During digitisation (also when outsourced), no new, separate rights arise in the case of a purely mechanical reproduction (book scanner, robotic scanner). This generally applies to mass digitisation of flatware. Furthermore, following the DSM Directive, it was clarified in Germany in the summer of 2021 that the reproduction of public domain visual works generally does not give rise to new property rights (cf. § 68 UrhG (German Copyright Act)).

If the digitised object itself it still protected by copyright, new, separate rights arise also in the case of reproduction photography. The same applies if a photograph is not a mere reproduction but has work character of its own. In the case of three-dimensional objects, demarcation becomes difficult as photographs may be considered photographic works in this context.

---

[89] http://www.persistent-identifier.de/.

[90] From a legal point of view, the starting point of copyright is always the work. Digitisations are reproductions of a work. Under licensing law, differences in format or resolution are irrelevant.

[91] https://openaccess.mpg.de/2365/en.

[92] DFG funding is not available for the processing of inaccessible material; a concept for making material accessible is a prerequisite, in terms of both the technical solution and timeframe.

If rights arise, it must be ensured that they accrue to the institution carrying out the project and that the institution, by using the release statement CC0, treats the reproductions as if these new rights did not exist. This is especially true for reproduction photography, but also for three-dimensional reproductions.

Generally, no rights arise from the creation of metadata with regard to the individual datasets.[93] If existing catalogue information is digitised, containing not only formal data but also descriptive text, and which therefore may be legally protected, any rights necessary for the provision and use of the data must be obtained.

In keeping with the principles of open access and open source, all results – metadata, digital images and full texts, or XML, DTDs or XML schema files underlying the full texts, as well as scripts and programming code – must be made available for reuse as freely as legally possible, either by marking them as public domain or by using free Creative Commons licences[94]or FLOSS licences for free software.

When digitising public-domain material, the digital copy must be marked as public domain. For all results where this is not possible because rights exist or have been generated in the process of digitisation, it is likewise necessary to ensure that the data are made available and licensed via open access as freely as legally possible. This can be done by completely waiving any ensuing rights with a CC0 mark (e.g. in the case of metadata) or by using the free licence types CC-BY or CC-BY-SA (→ 1.8). If protected material is digitised and the digital copy cannot be freely licensed because of rights to the source material that are still in effect, the legal status should be indicated with standardised rights statements.[95] For projects digitising not only public-domain materials and/or involving collaboration with commercial partners or publishers, delayed open-access publication (moving wall) of up to one year after the end of the project can be agreed.

6.1.2 Dealing with orphan works and unavailable works

Holdings of libraries, archives and museums often contain large quantities of out-of-print works that are of great value for research and education. However, clearing the rights of copyrighted materials proves to be extremely difficult, especially in the case of older materials and those that were not created in a utilisation context.

With the regulations on orphan works (§§ 61 et seq. UrhG (German Copyright Act)) and unavailable works (§§ 61d et seq. UrhG), legislators have therefore created legal permissions that enable cultural heritage institutions to put digitised works online – even without the necessary rights of use having been cleared. However, the accessibility thereby achieved does not meet the requirements for an open access publication, since, for example, Creative Commons licensing and thus subsequent use is not permitted. Permanent online availability is further not ensured, since in the case of both orphan and unavailable works, the rights holder can object to the online publication at any time. In addition, displaying the objects in the Deutsche Digitale Bibliothek (DDB, German Digital Library) or other portals is not possible even if it requires the transfer of rights of use to the portal. Where it is simply not legally permissible to make a work available online, it may be appropriate and necessary to proceed according to the regulations for orphan or unavailable works. Since this is an exception to the requirement of open access

[93] See Klimpel, Paul, Das Eigentum an Metadaten, in Euler, Ellen/Monika Hagedorn-Saupe/Gerald Maier/Werner Schweibenz/Jörn Sieglerschmidt; Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft, Berlin/Boston 2015, pp. 57 – 64.

[94] Always use the latest version of Creative Commons licences: https://creativecommons.org/.

[95] www.rightsstatements.org. See also the recommendations for rights information in metadata on the Wiki page of DINI AG KIM: https://wiki.dnb.de/pages/viewpage.action?pageId=217533652.

publication, its reasons must be stated, for example when the relevant material will become public domain in the near future anyway and the legal permissions are only needed for a short transition period.

In order to initiate a publication, first investigating the rights situation (identifying possible creators) and registering the works in relevant EU portals is required so that rights holders can find them.[96] Cultural heritage institutions are further subject to the following:

- removing images at the request of a rights holder,
- no provision of online download,
- no sharing to portals,
- no reuse of digital copies.

It is important to determine how long copyright protection will continue to apply and to ensure that the material is marked as public domain and made accessible without restrictions once the retention period ends.

In any case, it must be ensured that the metadata of such digitised material are permanently online so that the respective objects are available for use at the self-service points of the holding institution.

6.1.3 Access and reuse

The primary aim of making results available as freely as possible and using standardised rights statements is to create options for data not covered in section 6.1.2 and to be able to analyse digital images, metadata and full texts in usage contexts other than the immediate project environment (e.g. for data mining or data aggregation). This requires downloading, re-indexing and provision in separate research and presentation contexts. Digital images should be provided in a form that generally allows full scholarly use in other research contexts (e.g. through the use of individual images in frames, following the model of the DFG Viewer or the iFrame embedding by Google LLC or by issuing a general permission to display preview images on third-party servers). The data should be accessible without technical restrictions (login, IP address, etc.) and delivered in a way that allows CORS (Cross-Origin Resource Sharing). This is the only way to technically ensure that the data can actually be reused with other applications / web services without restriction. For this purpose, digitised images should be made available either as high-resolution derivatives in TIFF format or as fully resolved JPEG images with a compression of 90, maximum 80, in combination with the full scan resolution. This does not affect the levying of charges for copies in other qualities, derivatives or the production of other formats (CD, printouts and so forth).

The DFG expects that data made available online as part of DFG-funded projects include a clear reference to its origin and, if possible, a reference to DFG funding. In the case of digital images, this is still done by adding an acknowledgement to the published user copy on the website (derivative) or a footer line for the download of digital copies (for example in JPEG format).[97] This makes it easier to assign images to their origin later on (issue of provenance). Downloads of TIFF files should be provided without a footer. Manipulating the surface of digital images (e.g. by adding watermarks) is not permitted. For full texts an appropriate note may be added to the header of the text file; acknowledgements may also be incorporated in image headers. In any case, however, the

[96] https://euipo.europa.eu/ohimportal/en/web/observatory/orphan-works-db (orphan works); https://euipo.europa.eu/ohimportal/en/web/observatory/outofcommerceworks (out-of-commerce works). For further information on orphan works at the German Patent and Trade Mark Office, see https://www.dpma.de/english/our_office/index.html.

[97] A mandatory, visible proof of origin without visible alteration of the derivatives is currently being discussed. References regarding proof of origin could alternatively be included in the descriptive metadata, the image metadata and visually as part of the online presentation. However, this is not (yet) a recommendation of the DFG.

acknowledgement of origin and the funding credit must be included in the accompanying metadata and output in the data provision system.

6.2 Minimum requirements for deployment systems for digitised material

6.2.1 Functionality requirements

Collections / holdings can generally be made accessible in a variety of ways:

- via the website of the institution carrying out the project
  - all objects on the website are to be provided with navigation options
  - they must be published in a way that creates persistently citable URLs with the finest granularity possible. The proper citation format must be clearly indicated.
- via a locally implemented or externally operated DFG Viewer, if applicable to the material (→ 7.3)
- via references in the local and regional library catalogue / the local and regional archive portal
- via nationwide reference and presentation systems, especially the DDB or the Archivportal-D
- if available, via one of the DFG-funded material-specific portals that enable integrated access to all digital collections funded under the DFG programme, or a shared subject-specific portal, e.g. of the specialised information services
- via an OAI interface. The DFG Viewer too should be addressed via the OAI functionality (if the respective format is supported), whereby the interface has priority in any case.

Newly created OAI interfaces should be registered with the relevant portals and specialised information services. In addition, suitable measures should be taken to ensure that metadata can be found by search engines (e.g. with the sitemap protocol[98]).

As well as being able to access specific documents in a targeted way by means of a metadata search, users should also have the option of structured browsing in predefined collections, part-collections or holdings. Regarding the search engine it should be noted that simple, Google-style search tools are more intuitive to use than multi-field search masks that require a solid understanding of the data structure of a given collection or inventory. Ideally, both models should be combined in a faceted search, which enables users to whittle down long hit lists using defined criteria (facets). The highly specific multi-field search is still, however, desirable as an additional tool for a highly specialised community.

The following functions should also be implemented (provided no orphan or unavailable works are involved):

- Download function[99]
- Print function for output document view[100]
- Centralised DFG-funded information systems (VD16, VD17, VD18, subject-specific portals, material-specific portals, etc.) should first link to a view in the style of the DFG Viewer.

[98] http://www.sitemaps.org/.

[99] If the size of the complete file would be unmanageable, the file may be split into sections or individual pages for download purposes. Provision to researchers in the original size of the image master (at least JPEG 80% compression) should generally be made possible free of charge.

[100] If the size of the complete file would be unmanageable, the file may be split into sections or individual pages for print purposes.

6.2.2 Minimum technical requirements

As far as applicable, the presentation application / platform must be designed to:[101]

- Provide all materials in a quality that allows their convenient use for research purposes on typical university equipment. This entails, for instance, providing a type size that is easy to read.
- Provide all materials in a quality that allows processing without cumbersome delays.
- Enable the free download, for research purposes, of any complete unit as one single file (e.g. of individual printed works).
- Support all currently popular browsers, to the extent viable.[102]

7. Presentation Standards for Text-Based Works, DFG Viewer

The principles laid out above apply to any kind of project that provides digital content. Additionally, the following minimum requirements apply to the provision of digital files that have the character of digital books or multi-page documents. They cover certain basic standards and the minimum with regard to functionality.

7.1 Basic requirements and architecture

The provision system combines digitised image or full-text files into a document structure to enable users to navigate a document. Furthermore, it establishes connections between digital documents, or parts thereof (e.g. chapters, pages), and metadata, to allow users to access the individual document or certain document parts based on a metadata search. Finally, it organises digital documents into digital collections or holdings according to subject matter or origin, to let users navigate documents and collections in a scientifically appropriate fashion. It provides user interfaces for searching, navigating, accessing and retrieving metadata, documents, collections and holdings, and it supports largely automated export and import of standards-compliant raw data. The provision systems of the individual information infrastructure facilities should allow access across institutions, both in navigating digital collections or holdings and in searching indexes. In addition, the transparent linkage of provision systems with local catalogue / information systems as well as with overarching information systems is desirable.

Various system architectures can be used to accomplish these tasks. The following basic alternatives are viable:

- Metadata are stored centrally in an online information system (e.g. the local OPAC or online search system, or an overarching information system such as a library network catalogue or Archivportal-D), while digital document files including XML-encoded structure data are provided on a separate document server for online access. The structure of the digitised collection or the internal structure of the digitised documents can be mirrored using the hierarchy of the file system.
- A document management system (DMS) or Content Management System (CMS) is used, in which both the metadata and the digital data are stored in the database system.

[101] The key criterion is practicality, not the implementation of abstract desirables. If objects in a project by their nature cannot be meaningfully displayed with a resolution under 1600 × 1200, there is no need to bother with pseudo solutions; if an object cannot be processed meaningfully under 3 MB, it does not violate the criterion of DSL compatibility not to provide smaller versions.

[102] If a browser does not support a format required by an advanced 3D application, there is no need to make the effort of developing a suitable plug-in.

Typically, the first variant is used, which allows for a distributed and transparent information infrastructure.

7.2 Functionality requirements

A key benchmark of functional quality is the comfort with which users can navigate within a found document. The following navigation functions are considered the basic standard:

- Go to any desired image
- Home: Jump to beginning of document
- End: Jump to end of document
- Forward: Go forward one page
- Back: Go back one page
- Full text search within digital copies, if available (mandatory for books from 1850 onward)[103]
- Metadata info: Here users can view current document information in description fields stored in the information system.
- Help: Help menu should provide detailed descriptions with examples for navigation and for searching the digital library.

Whenever possible and appropriate, tables of contents, structure trees or functional equivalents should be included and designed to be searchable. Navigation aids are desirable, e.g. graphic representations in a header that indicate to the user the current location within the digital document. If an information system contains materials that users will normally regard as conceptual units (multivolume works), these units must be visible as such.

7.3 DFG Viewer and IIIF

In addition to the differently designed and locally managed web information systems of individual institutions, scientific users should have standardised access to the data (contents) of all DFG-funded digitised works. To this end, the DFG currently pursues two complementary strategies:

- Defining a standardised design profile for visualising digital copies that were generated with DFG funding ("DFG Viewer"), thus enabling uniform presentation of all DFG-funded projects. At the same time, the viewer serves as a reference implementation for the required standards, which institutions carrying out digitisation can use to check whether they meet these standards (verification of the desired implementation). The DFG Viewer will continue to provide this function.
- More important than integrating the viewer is the creation of an OAI interface based on the METS standard: Its primary purpose is to display derivatives and their metadata in a uniform manner for all DFG-funded projects and, depending on material, to deliver METS/MODS for printed works and archive materials, and METS/TEI for manuscripts.

For their primary transregional presentations, DFG-funded digitisation projects should use the aforementioned interfaces. However, it is not mandatory to implement the DFG Viewer at one's own institution.[104] Instead, other viewers with a comparable range of functionality may be used. Appendix A defines the METS/MODS format and Appendix B the METS/TEI format of the interface as well as the viewer design.

[103] With today's technology, OCR should always be considered for machine-press era printed works from 1850 onward.

[104] A local installation by the aggregating portal or the data-supplying institution is not necessary in this case.

As the International Image Interoperability Framework (IIIF), which unites several APIs, is becoming increasingly common, the IIIF Presentation API should also be referred to here as an equivalent option for data provision. IIIF guarantees a high level of interoperability when exchanging digital objects. However, it cannot replace the above-mentioned reference status of the DFG Viewer standard, which is why it is recommended as a priority by the current Practical Guidelines.

For the following appendices A-C, please note the documentation on the formats for delivering data to the DDB:

https://wiki.deutsche-digitale-bibliothek.de/display/DFD/Lieferformate

Appendices

Appendix A: METS/MODS Profile for DFG Viewer Display and Transmission by OAI

1. DFG Viewer

In order to achieve a uniform presentation when local digital offerings are accessed via nationwide catalogue systems (e.g. Deutsche Digitale Bibliothek, Archivportal-D), DFG-funded projects should use the browser display known as DFG Viewer as well as the interfaces on which it is based. The purpose is to make it easier for researchers to use digitised content. The DFG Viewer may then link to the specific local information systems of any given institution.

The DFG Viewer[105] was developed in the context of the funding scheme "Digitisation of VD16 / VD17" by the libraries funded in the first round of proposals. The SLUB Dresden, in collaboration with additional partners, continues to develop the Viewer on an ongoing basis. The Viewer's reference application is currently hosted by the SLUB Dresden.

To give DFG-funded projects maximum certainty for proposal planning and ensure that metadata meets the DFG Viewer requirements, the metadata generated by such projects should be valid against one of the application profiles on the Viewer's website.[106]

METS[107] is used to display metadata in the DFG Viewer. It serves as a frame (wrapper) within which descriptive, administrative and structural metadata as well as resources (e.g. digital images, full texts) are recorded. To display descriptive metadata, the DFG Viewer requires MODS[108] encoded metadata (for prints, journals, newspapers, archival materials) or TEI encoded metadata (for medieval and early modern manuscripts). For administrative metadata (e.g. rights information and link to the local application), a special format developed for the Viewer (Namespace dv) is used.

Detailed documentation on how to implement the METS format can be found on the website of the DFG Viewer's reference application.

These guidelines apply in principle to all digitised text-based works. For medieval and early modern manuscripts, METS/TEI can be used as an alternative format or in addition (→ Appendix B). For pictorial and three-dimensional objects, using LIDO instead of METS/MODS or METS/TEI is mandatory (→ Appendix C). EAD(DDB) must be used to encode archival finding aids.[109]

[105] https://dfg-viewer.de/.

[106] https://dfg-viewer.de/metadaten.

[107] http://www.loc.gov/standards/mets/.

[108] https://www.loc.gov/standards/mods/.

[109] See https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=19010180.

2. MODS DFG standard set (library holdings)

For DFG Viewer display, only a few mandatory fields are required (see tables below).

1 Title information

Element /subelement	Repeatable	Comments	Status
<titleInfo>	Yes	Title information; if work has no title, a meaningful title must be created. In the case of volumes or issues of multivolume works without individual titles for each volume, journals or newspapers, it is sufficient to state the title of the entire work in the element <mods:relatedItem type="host">. However, it is then mandatory to state the numbering in the element <mods:part> with the subelements <mods:detail> and <mods:number>.	Mandatory for monographs and for volumes of multivolume works with individual titles
<titleInfo> / <title>	No	Contains main title of work.	Mandatory
<titleInfo> / <subTitle>	Yes	Contains subtitle / addition to main title of work.	Mandatory if applicable

See examples of mods:titleInfo https://wiki.deutsche-digitale-bibliothek.de/display/DFD/titleInfo

2 Person

Element /subelement	Repeatable	Comments	Status
<name type="personal" valueURI="...">	Yes	Person related to work (e.g. author). The attribute @valueURI contains the URI of the person or of the person's name. If the @valueURI is not an HTTP URI, @authority must contain the code for the set of rules.	Mandatory if applicable
<mods:part>	Yes	Contains name elements of the type specified in @type; possible values are "date", "family", "given", "termsOfAddress".	
<name> / <displayForm>	No	Name in desired display form	Mandatory if mods:name applies
<name> / <role>	Yes	Wrapper element for role of person	Mandatory if mods:name applies
<name> / <role> / <roleTerm type="code" authority="marcrelator">	No	Role of person; the value in mods:roleTerm must be a code from the MARC Relator vocabulary (https://id.loc.gov/vocabulary/relators.html)	Mandatory if mods:name applies

See examples https://wiki.deutsche-digitale-bibliothek.de/display/DFD/name

3 Corporate body

Element /subelement	Repeatable	Comments	Status
<name type="corporate" authority="...">	Yes	Corporate body related to work.	Mandatory if applicable
<name> / <namePart> <name> / <displayForm>	Yes No	See above	 Mandatory if mods:name applies
<name> / <role>	No	See above	Mandatory if mods:name applies.
<name> / <role> /	No	See above	
<roleTerm type="code" authority="marcrelator">			

See examples of mods:name https://wiki.deutsche-digitale-bibliothek.de/display/DFD/name

4 Publication information / imprint

Element /subelement	Repeatable	Comments	Status
<originInfo eventType="publication">	Yes	The first <originInfo> block is for information on the source; the second <originInfo> block is for including information on the digital edition by indicating eventType="digitisation".	Mandatory if applicable
<originInfo>/<place>/ <placeTerm type="text">		Contains name of place of publication.	Mandatory if applicable
<originInfo> / <publisher>	Yes	Contains name of publisher or print shop.	Mandatory if applicable
<originInfo> / <dateIssued encoding="8601">	Only if attribute @point is used	Contains year of publication.	Mandatory if applicable

See examples of mods:originInfo https://wiki.deutsche-digitale-bibliothek.de/display/DFD/originInfo

5 Edition information

Element /subelement	Repeatable	Comments	Status
<originInfo> / <edition>	Yes	Contains name of edition (see also above, information about electronic edition).	Mandatory if applicable

6 Physical description

Element /subelement	Repeatable	Comments	Status
<physicalDescription>	No	Physical description area / collation	Mandatory
<physicalDescription> / <extent>			

See examples of mods:physicalDescription https://wiki.deutsche-digitale-bibliothek.de/display/DFD/physicalDescription

7 Superior work level

Element /subelement	Repeatable	Comments	Status
<relatedItem type="host"> / <recordInfo> / <recordIdentifier @source>	No	<recordIdentifier> within mods:relatedItem type=”host” is an identifier that permits linkage to hierarchically superior / superordinate datasets. The @source attribute must contain information on origin of dataset and is mandatory.	Mandatory if hierarchy applies

See examples of mods:relatedItem https://wiki.deutsche-digitale-bibliothek.de/display/DFD/relatedItem

8 Volume information

Element /subelement	Repeatable	Comments	Status
<part order="..."> / <detail> / <number>	No	Part information; value of @order attribute is any numeric value (integer) that ensures correct order of parts; <number> states numbers of volumes, issues, etc.	Mandatory if applicable

See examples of mods:part https://wiki.deutsche-digitale-bibliothek.de/display/DFD/part

9 Language

Element /subelement	Repeatable	Comments	Status
<language>	Yes	Language information	Mandatory
<language> / <languageTerm type="code" authority="iso639-2b">	No	Contains language of work in ISO 639-2/B code.	

See examples of mods:language https://wiki.deutsche-digitale-bibliothek.de/display/DFD/language

10 Citable identifier

Element /subelement	Repeatable	Comments	Status
<identifier type="...">	Yes	Globally unique identifier of resource (@type attribute, e.g. URN, PURL, DOI, Handle, URI, etc.). If available, GW and VD numbers of print must also be indicated (@type attributes VD16, VD17, GW).	Mandatory

See examples of mods:identifier https://wiki.deutsche-digitale-bibliothek.de/display/DFD/identifier

11 Database ID

Element /subelement	Repeatable	Comments	Status
<recordInfo> / <recordIdentifier @source=>	No	Dataset identifier for unique identification within a database system, e.g. PICA production number. The @source attribute must contain information on origin of dataset and is mandatory.	Mandatory

See examples of mods:recordInfo https://wiki.deutsche-digitale-bibliothek.de/display/DFD/recordInfo

12 Shelfmark

Element /subelement	Repeatable	Comments	Status
<location> <physicalLocation> <shelfLocator>	Yes No Yes	Location and call number of original	Recommended

See examples of mods:location https://wiki.deutsche-digitale-bibliothek.de/display/DFD/location

Depending on the type of material and design of the project, an expansion of this basic set should be considered. The DFG Viewer website offers more differentiated format definitions for this purpose, including additional explanations of how to use the various fields.

The requirements for delivering METS/MODS data to the DDB are defined in the DDBinfo Wiki.[110] It also contains a link to the Schematron validation of the DDB.[111]

[110] https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=19006651.

[111] https://wiki.deutsche-digitale-bibliothek.de/display/DFD/Schematron-Validierungen+der+Fachstelle+Bibliothek#SchematronValidierungenderFachstelleBibliothek-METS/MODS-Anwendungsprofilf%C3%BCrdigitalisierteMedien.

3. Notes for archives on converting EAD(DDB) to MODS

Mapping EAD(DDB) to corresponding MODS elements is generally easy. For simplified processing of archival datasets, tools are available in the archiving community for use by interested institutions.[112]

Field description	EAD(DDB)	MODS	Status
Title	c/did/unittitle	mods:titleInfo / mods:title and further subelements if appropriate MODS definition: A word, phrase, character, or group of characters, normally appearing in a resource, that names it or the work contained in it. Comment: For the root structure element in a METS/MODS set, at least one title must be specified using the mods:titleInfo element with the mods:title subelement.	Mandatory
Institution	ancestor::archdesc/did/repository/corpname	mods:location / mods:physicalLocation MODS definition: The institution or repository that holds the resource or where it is available.	Mandatory
Shelfmark	c/did/unitid[not(@type)]	mods:location / mods:shelfLocator MODS definition: Shelfmark or other shelving designation that indicates the location identifier for a copy.	Mandatory
Duration	c/did/unitdate	mods:originInfo / mods:dateCreated MODS definition: The date of creation of the resource Comment: Date of creation or duration of the analogue source material.	Mandatory if applicable

[112] For this purpose, the DDB Archives Department provides the Data Preparation Tool: https://wiki.deutsche-digitale-bibliothek.de/display/DFD/DDB+Data+Preparation+Tool, accessed on October 25, 2022. This application is being further developed and maintained by the Archives department.

Field description	EAD(DDB)	MODS	Status
Context[113]	ancestor::c/did/unit-title	mods:relatedItem type="host" / mods:title if appropriate with further subelements MODS definition: Information that identifies other resources related to the one being described. mods:relatedItem can be used to show not only hierarchical relationships, but a variety of relationships using the @type attribute. To create a relationship between documents or document parts with MODS, unique identifiers must be used.	Recommended
ID	c/@id	mods:recordInfo / mods:recordIdentifier MODS definition: Contains the system control number assigned by the organization creating, using, or distributing the record.	Mandatory

4. Examples of DFG Viewer-compatible METS/MODS datasets

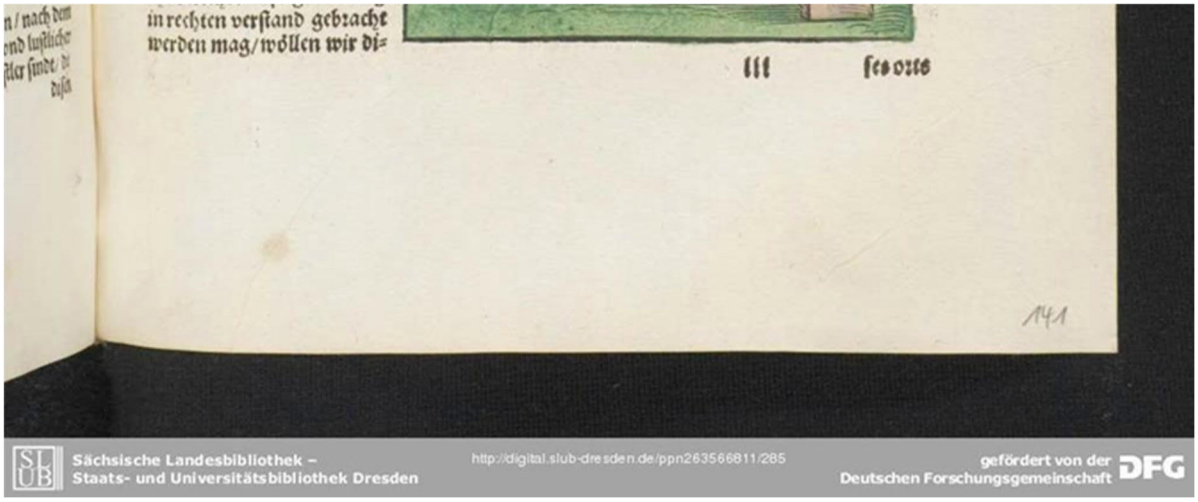
(1) https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=19006655

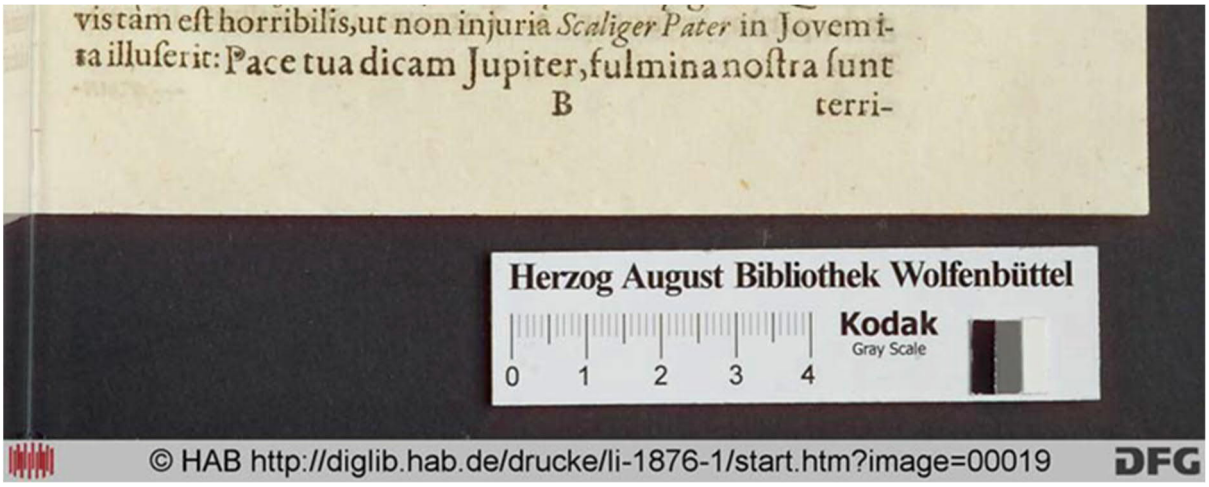
(2) https://wiki.deutsche-digitale-bibliothek.de/display/DFD/5.+Beispiele+und+Schemata+zur+Validierung#id-5.BeispieleundSchematazurValidierung-Beispieldateienf%C3%BCrdasarchivischeMETS/MODS-Profil

5. Adding a footer

For formats provided to users or on the Internet, the footer should be appended to the lower edge of the digital image. The institution logo should be shown on the left side of the footer; for DFG-funded projects, the DFG logo should be added on the right side, if possible. It is recommended to provide a citable URL in the middle area (in addition to including it in the Viewer XML file). Fonts and logos must be scaled according to the resolution. In PDF files for download, footers may also be added to each image, in addition to a cover page. See the following examples: Mapping EAD(DDB) to corresponding MODS elements is generally easy to achieve.

[113] This refers to the inventory context of the resource, i.e. the higher levels of description in terms of tectonics and classification.


in rechten verſtand gebracht
werden mag/wöllen wir di=
fes ortes
141
Sächsische Landesbibliothek –
Staats- und Universitätsbibliothek Dresden
http://digital.slub-dresden.de/ppn263566811/285
gefördert von der
Deutschen Forschungsgemeinschaft
DFG


vis tàm eſt horribilis, ut non injuria Scaliger Pater in Jovem i-
ſa illuſerit: Pace tua dicam Jupiter, fulmina noſtra ſunt
B terri-
Herzog August Bibliothek Wolfenbüttel
Kodak
Gray Scale
0 1 2 3 4
© HAB http://diglib.hab.de/drucke/li-1876-1/start.htm?image=00019
DFG


Leipzig / Bey Timotheo Ritzſchen.
Universitäts- und Landesbibliothek
Sachsen-Anhalt
urn:nbn:de:gbv:3:1-90640-p0001-4
DFG

Appendix B: METS/TEI Specification for the Display of Digitised Manuscripts

1. DFG Viewer

The objective of this METS/TEI specification for digitised medieval and early-modern manuscripts is the display of the digitised material within the unified interface of the DFG Viewer. The DFG Viewer, which was originally created to display digitised printed works, is thus extended to another media type without abandoning the familiar user interface. It is merely supplemented by a few extra functions to meet the specific requirements of the new media type.

Since this is not a fundamentally new implementation of the DFG viewer, but merely the support of an additional application profile for METS/TEI, the information given in Appendix A on printed works also applies to digitised manuscripts – in particular the general information on the DFG Viewer in section 1, as well as the instructions for adding a footer in section 5.

More information on the data formats and functions supported by the DFG Viewer as well as on how to integrate the web service in your own information systems can be found on the DFG Viewer website. [114]

2. TEI-DFG standard set (manuscript holdings)

The following information on how to encode a digitised manuscript in TEI is limited to the necessary mandatory fields and does not cover the full spectrum of possibilities. Detailed documentation of all options can be found on the DFG Viewer website. The basis of the application profile is the TEI specification of the Text Encoding Initiative[115] as well as the preparatory work of the Europeana Regia project,[116] with which the DFG standard set is compatible.

1 Identification data

All elements described in this section are subelements of teiHeader/fileDesc/sourceDesc/msDesc

Element /subelement	Repeatable	Comments	Status
<head> / <title>	No	Contains the title of the manuscript according to the DFG Guidelines for Cataloguing Manuscripts.	Mandatory if applicable
<msIdentifier> / <settlement>	No	Contains the geographical location where the manuscript is held.	Mandatory
<msIdentifier> / <repository>	No	Contains the name of the institution holding the manuscript.	Mandatory
<msIdentifier> / <idno>	No	Contains an identifier that is unique within the institution holding the manuscript, usually the shelfmark, supplemented, where appropriate, by an indication of the folios across which the logical structure extends.	Mandatory
<msIdentifier> / <msName>	No	Contains a non-canonical title of the manuscript.	Recommended

[114] https://dfg-viewer.de/.

[115] https://tei-c.org/.

[116] http://www.europeanaregia.eu/.

2 Information on the origin
All elements described in this section are subelements of
teiHeader/fileDesc/sourceDesc/msDesc/history

Element /subelement	Repeatable	Comments	Status
<origin> / <origPlace>	No	Contains the place of origin of the manuscript as it should appear in the DFG Viewer.	Recommended
<origin> / <origDate>	No	Contains information about the time of origin of the manuscript as it should appear in the DFG Viewer. Exact numeric specifications are to be made in the attributes @when, @notBefore, @notAfter, @from and @to, if possible.	Recommended

3 Information on the physical condition
All elements described in this section are subelements of
teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc

Element /subelement	Repeatable	Comments	Status
<supportDesc material="...">			Mandatory if applicable
<supportDesc> / <support>	No	Contains information on the writing material as it should appear in the DFG Viewer.	Mandatory if applicable
<supportDesc> / <extent> / <measure>	No	Contains quantitative information regarding volume, such as the number of folios, which is also displayed in the DFG viewer.	Mandatory if applicable
<supportDesc> / <extent> /<dimensions>	No	Contains information on physical dimensions, such as the sheet format, which is also displayed in the DFG Viewer. The precise size specifications are to be made in the subelements depth, height and width.	Mandatory if applicable

4 Information on content
All elements described in this section are subelements of
teiHeader/fileDesc/sourceDesc/msDesc/msContent

Element /subelement	Repeatable	Comments	Status
<summary>	No	Contains a summary description of the content.	Recommended
<textLang mainLang="..." otherLangs="...">	No	Contains information on the languages in which the texts of the manuscript are written. The @mainLang attribute should specify the language predominantly used, whereas all other occurring languages should be encoded in the @otherLangs attribute separated by spaces. The attribute values must be language codes according to the standards ISO 639-2 or ISO 639-3.	Recommended

5 Administrative information on the digital copy
All elements described in this section are subelements of teiHeader/fileDesc

Element /subelement	Repeatable	Comments	Status
<titleStmt> / <title>	No	Contains a title for the digital copy. This is typically the headline, consisting of shelfmark and title of the digitised manuscript, according to DFG guidelines. This field should not be confused with the title and is not displayed in the DFG Viewer. The requirement is based on the general TEI specification and is not specific to the DFG Viewer.	
<titleStmt> / <funder>	Yes	Contains a funding body in connection with the digitisation, e.g. “Deutsche Forschungsgemeinschaft“. It may also be the institution carrying out the digitisation.	Recommended
<publicationStmt> / <publisher>	Yes	Contains the publisher of the digital copy, typically the institution carrying out the digitisation. At least one subelement of publicationStmt is mandatory.	Mandatory if applicable
<publicationStmt> / <pubPlace> / <ptr target="..." cRef="..." type="..." />	Yes	Contains a persistent identifier of the digital copy. If the identifier is a URL, it is to be specified in the @target attribute, otherwise in the @cRef attribute. In addition, the type of identifier must be indicated using the values urn or purl in the @type attribute. At least one subelement of publicationStmt is mandatory.	Mandatory if applicable
<publicationStmt> / <idno>	Yes	Contains an identifier of the dataset. Optionally, its type may be specified in the @type attribute.	Mandatory if applicable

3. Example of a METS/TEI dataset according to the DFG standard

Due to its length, the following example shows only an excerpt of the METS/TEI dataset. The excerpt represents a complete descriptive metadata record, exemplarily containing all mandatory and optional fields described above. This extract is located within a dmdSec section of a METS file.

```
<tei:teiHeader>
  <tei:fileDesc>
    <tei:titleStmt>
     <!-Fingierter Titel für das Digitalisat (obligatorisch)
     Es handelt sich hierbei nicht um den bibliografischen Titel der Handschrift, sondern um die nach DFG-
     Richtlinien gebildete Schlagzeile bestehend aus Signatur und wissenschaftlichem Titel
     Der Titel kommt im DFG-Viewer nicht zur Anzeige, ist aber aufgrund der TEI Spezifikation dennoch
     verpflichtend. -->
      <tei:title>
        Zeugbuch Kaiser Maximilians I. BSB Cod.icon. 222
      </tei:title>
     <!-Förderer der Digitalisierung (optional) -->
      <tei:funder>Deutsche Forschungsgemeinschaft</tei:funder>
    </tei:titleStmt>
   <tei:publicationStmt>
     <!-Herausgeber des Digitalisats (obligatorisch, sofern bekannt) -->
      <tei:publisher>Bayerische Staatsbibliothek</tei:publisher>
     <!-Persistenter Identifikator für das Digitalisat (obligatorisch)->
      <tei:pubPlace>
        <tei:ptr
        target="http://daten.digitale-sammlungen.de/~db/0002/bsb00020956/"
        type="purl"/>
     </tei:pubPlace>
     <!—Identifikator für den Datensatz (obligatorisch) -->
      <tei:idno type="zend">bsb00020956</tei:idno>
     <!-Lizenzangabe->
      <tei:availability>
        <tei:licence
        target="http://creativecommons.org/publicdomain/zero/1.0/"/>Public domain (CC0)
         </tei:licence>[117]
     </tei:availability>
    </tei:publicationStmt>
   <tei:sourceDesc>
     <tei:msDesc>
       <tei:msIdentifier>
         <!-Geographischer Aufbewahrungsorts (obligatorisch) -->
         <tei:settlement>München</tei:settlement>
         <!-Name der bewahrenden Institution (obligatorisch) -->
         <tei:repository>Bayerische Staatsbibliothek</tei:repository>
         <!-Signatur o. eindeutiger Identifikator (obligatorisch) -->
         <tei:idno>Cod.icon. 222</tei:idno>
        <!-Nicht-kanonischer Titel (optional, wiederholbar) -->
          <tei:msName>Zeugbuch Kaiser Maximilians I.</tei:msName>
       </tei:msIdentifier>
       <tei:head>
         <!-Handschriftentitel (obligatorisch, sofern vorhanden) -->
          <tei:title/>
       </tei:head>
       <tei:history>
         <tei:origin>
```

[117] If required for the DFG Viewer, the same information must be added again in dv:licence.

```
            <!-Entstehungsort (empfohlen, sofern bekannt) -->
            <tei:origPlace>Innsbruck</tei:origPlace>
            <!-Entstehungszeit (empfohlen, sofern bekannt) -->
            <tei:origDate notBefore="1500" notAfter="1505">
              um 1502
            </tei:origDate>
          </tei:origin>
        </tei:history>
        <tei:physDesc>
          <tei:objectDesc>
            <!-Angabe des Materials (obligatorisch, sofern ermittelbar) -->
            <tei:supportDesc material="mixed">
              <!-Angaben zum Beschreibstoff (obligatorisch, sofern
                ermittelbar) -->
              <tei:support>
                Pergament (2-8, 11-30, 34, 48, 295, 296) und Papier
              </tei:support>
              <tei:extent>
                <!-Angaben zum Umfang wie z.B. die Blattzahl
                  (obligatorisch, sofern ermittelbar) -->
                <tei:measure type="leavesCount">2 + 276 Blätter</tei:measure>
                <!-Blattformat (obligatorisch, sofern ermittelbar) -->
                <tei:dimensions unit="mm" type="leaves">
                  <tei:height>420</tei:height>
                  <tei:width>285</tei:width>
                </tei:dimensions>
              </tei:extent>
            </tei:supportDesc>
          </tei:objectDesc>
        </tei:physDesc>
        <tei:msContents>
          <!-Summarische Beschreibung des Inhalts (empfohlen) -->
          <tei:summary/>
          <!-Textsprache (empfohlen, sofern bekannt) -->
          <tei:textLang mainLang="ger">Deutsch</tei:textLang>
        </tei:msContents>
      </tei:msDesc>
    </tei:sourceDesc>
  </tei:fileDesc>
</tei:teiHeader>
```

The remaining structure of the METS file is analogous to the documentation for printed works. The logical structure of the manuscript is encoded in the METS file within the structMap element with the TYPE=LOGICAL attribute. Each logical structural unit (e.g. a text fragment) may in turn refer to its own section with descriptive metadata, which must also follow the structure described above.

Appendix C: LIDO core elements for publication

1. LIDO core elements for pictorial and three-dimensional material

The LIDO standard enables both sharing and provision of descriptive and administrative metadata for searching and presenting rare, pictorial and three-dimensional objects in online environments, thus supporting the joint use and linking of data on the Internet. Further information on LIDO can be found on the LIDO website at http://www.lido-schema.org.

LIDO defines only a few mandatory elements. In addition to these mandatory LIDO elements highlighted in red, the table below contains further LIDO elements that are useful for object documentation in digitisation projects. For elements marked with "Mandatory if applicable", there may be reasons why they cannot be assigned, e.g. because the information is not useful for the respective object type. Here, specifications may be adjusted for the specific use case. Depending on project context, material-specific or subject-specific LIDO profiles may be used for orientation when applying LIDO: The LIDO Handbuch Graphik[118] should be consulted for recording and publishing metadata on drawings and prints, and the LIDO Handbuch Malerei und Skulptur[119] for paintings and sculptures. Further information about LIDO profiles are available via the thematic portal "LIDO for cultural objects"[120] and the LIDO website[121]. Apart from the general LIDO mandatory elements, LIDO profiles may contain further profile-specific mandatory elements and provide information on other recommended elements.

For further information on the LIDO elements listed below and others, please refer to the documentation of the latest LIDO version.[122] It contains a description of all LIDO elements according to the following model: https://lido-schema.org/schema/latest/lido.html#objectWorkType

Many LIDO elements (e.g. object type, material, technique) are assigned data values that should ideally be defined as terms ("units of thought"[123]) in controlled vocabularies (thesauri, classification systems or word lists).[124] Such elements are indicated by the word Term in the respective comments section below. To reference the controlled vocabulary, these LIDO elements have the subelements defined in http://lido-schema.org/schema/latest/lido.html#conceptComplexType. Some LIDO elements (e.g. type of attribution) may preferably be assigned as a term, but alternatively also with free text, following the definition at http://lido-schema.org/schema/latest/lido.html#conceptMixedComplexType. The document "LIDO Terminology Recommendation"[125] provides a wide range of recommendations on the use of LIDO terminology[126] and on the use of other vocabularies.

Recommendations from the LIDO Terminology Recommendation Document for the respective element (if available) can be retrieved according to the following pattern:
https://lido-schema.org/documents/terminology-recommendation.html#objectWorkType

[118] https://doi.org/10.11588/arthistoricum.382.544.

[119] https://doi.org/10.11588/arthistoricum.1026.

[120] https://www.arthistoricum.net/themen/portale/lido.

[121] http://www.lido-schema.org.

[122] https://lido-schema.org/schema/latest/lido.html.

[123] https://www.w3.org/TR/2009/REC-skos-reference-20090818/#concepts.

[124] The vocabularies may (preferably) be published or unpublished and locally maintained.

[125] http://lido-schema.org/documents/terminology-recommendation.html.

[126] http://terminology-view.lido-schema.org/.

1 Identifiers for LIDO dataset and object

Element /subelement	Repeatable	Comments	Status
<lidoRecID>	Yes	https://lido-schema.org/schema/latest/lido.html#lidoRecID Unique identification number for the LIDO dataset, typically comprising the data supplier's ISIL number and the local dataset number. (See https://sigel.staatsbibliothek-berlin.de/, for ISIL numbers for museums see https://isil.museum/)	Mandatory
<objectPublishedID>	Yes	https://lido-schema.org/schema/latest/lido.html#objectPublishedID A published identifier of the object or work, preferably in the form of a dereferenceable URI.	Mandatory if applicable

Descriptive metadata

2 Object types

Element /subelement	Repeatable	Comments	Status
<objectWorkType>	Yes	https://lido-schema.org/schema/latest/lido.html#objectWorkType Object or work type Term	Mandatory

3 Classification

Element /subelement	Repeatable	Comments	Status
<classification type="...">	Yes	https://lido-schema.org/schema/latest/lido.html#classification A term that categorises an object within a larger context. The (optional) @type attribute qualifies the type of qualification; for possible values, see http://terminology.lido-schema.org/classification_type	Recommended

4 Title / object name

Element /subelement	Repeatable	Comments	Status
<titleSet>	Yes	https://lido-schema.org/schema/latest/lido.html#titleSet Wrapper for title information	Mandatory
<titleSet> / <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#appellationValue Title or name given to an object. If the object has no title it must be generated, e.g. from the object type and an identifier, such as the inventory number.	Mandatory

5 Repository / location

Element /subelement	Repeatable	Comments	Status
<repositorySet type=" http://terminology.lido-schema.org/lido00475">	Yes	https://lido-schema.org/schema/latest/lido.html#repositorySet Wrapper for information about current repository / location; can contain the type attribute =“http://terminology.lido-schema.org/lido00476“ to include additional information on former repositories / locations.	Mandatory if applicable
<repositorySet> / <repositoryName> / <legalBodyName> / <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#repositoryName Name of institution holding the object	Mandatory if applicable
<repositorySet> / <workID>	Yes	https://lido-schema.org/schema/latest/lido.html#workID A unique numerical or alphanumerical identification number given to the object by the holding institution; typically the inventory number	Mandatory if applicable
<repositorySet> / <repositoryLocation> / <placeID>	Yes	https://lido-schema.org/schema/latest/lido.html#placeID Identifier for the location. Ideally references an authority file.	Recommended
<repositorySet> / <repositoryLocation> / <namePlaceSet> / <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#repositoryLocation Name of location (mandatory in particular for architectural objects)	Mandatory if applicable

6 Object description

Element /subelement	Repeatable	Comments	Status
<objectDescriptionSet type="...">	Yes	https://lido-schema.org/schema/latest/lido.html#objectDescriptionSet Wrapper for descriptive texts and their sources. The (optional) @type attribute qualifies the type of description; for possible values, see http://terminology.lido-schema.org/objectDescriptionSet_type	Recommended
<objectDescriptionSet> / <descriptiveNoteValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#descriptiveNoteValue	Recommended
<objectDescriptionSet>/ <objectDescriptionRights>	Yes	https://lido-schema.org/schema/latest/lido.html#objectDescriptionRights Rights to the descriptive text	Recommended

7 Measurements

Element /subelement	Repeatable	Comments	Status
<objectMeasurementsSet>	Yes	https://lido-schema.org/schema/latest/lido.html#objectMeasurementsSet Wrapper for measurements The specification of measurements is also possible in relation to events in <eventSet> / <event> / <eventObjectMeasurements>. See https://lido-schema.org/schema/latest/lido.html#eventObjectMeasurements	Mandatory if applicable
<objectMeasurementsSet> / <displayObjectMeasurements>	For language variants only	https://lido-schema.org/schema/latest/lido.html#displayObjectMeasurements Measurements given in text form	Mandatory if applicable
<objectMeasurementsSet> / <objectMeasurements> / <measurementsSet>	Yes	https://lido-schema.org/schema/latest/lido.html#measurementsSet Measurements given in structured form with the subelements <measurementType>, <measurementUnit>, <measurementValue>	Recommended

8 Event

Element /subelement	Repeatable	Comments	Status
<eventSet>	Yes	https://lido-schema.org/schema/latest/lido.html#eventSet Wrapper for information about an event in the object's life cycle	Mandatory if applicable
<eventSet> / <event> / <eventType>	No	https://lido-schema.org/schema/latest/lido.html#eventType Type of event Depending on the object type, typical events might be: production, discovery, use. For recommended event types, see http://terminology.lido-schema.org/eventType Term	Mandatory if applicable
Identifiable actors involved in the event			
<eventSet> / <event> / <eventActor>	Yes	https://lido-schema.org/schema/latest/lido.html#eventActor Wrapper for information about the involvement of a person, institution, group or family in the event	Mandatory if applicable
<eventSet> / <event> / <eventActor> / <actorInRole> / <actor> / <actorID>	Yes	https://lido-schema.org/schema/latest/lido.html#actorID Identifier for the actor involved. Ideally references an authority file. GND should be used for DFG projects.	Mandatory if applicable
<eventSet> / <event> / <eventActor> / <actorInRole> / <actor> / <nameActorSet> / <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#actorInRole https://lido-schema.org/schema/latest/lido.html#appellationValue Name of actor involved	Mandatory if applicable
<eventSet> / <event> / <eventActor> / <actorInRole> / <roleActor>	Yes	https://lido-schema.org/schema/latest/lido.html#roleActor Role of actor involved Term	Recommended
<eventSet> / <event> / <eventActor> / <actorInRole> / <attributionQualifierActor>	Yes	https://lido-schema.org/schema/latest/lido.html#attributionQualifierActor Information for attributing an object to an actor; particularly relevant for artists	Recommended

Element /subelement	Repeatable	Comments	Status
Cultural context of event			
<eventSet> / <event> / <eventActor> / <actorInRole> / <attributionQualifierActor>	Yes	https://lido-schema.org/schema/latest/lido.html#attributionQualifierActor Information for attributing an object to an actor; particularly relevant for artists	Recommended
Date of event			
<eventDate>	No	https://lido-schema.org/schema/latest/lido.html#eventDate Date of event	Mandatory if applicable
<eventDate > / <displayDate>	For language variants only	https://lido-schema.org/schema/latest/lido.html#displayDate Date in text form, which allows any uncertainties to be noted.	Mandatory if applicable
<eventDate > / <earliestDate>	No	https://lido-schema.org/schema/latest/lido.html#earliestDate Year or exact date representing earliest date on which the event took place or started.	Mandatory if applicable
<eventDate > / <latestDate>	No	https://lido-schema.org/schema/latest/lido.html#latestDate Year or exact date representing latest date on which the event took place or ended.	Mandatory if applicable
Time period of event			
<periodName>	Yes	https://lido-schema.org/schema/latest/lido.html#periodName Time period of event; e.g. archaeological, historical or artistic period. Term	Recommended

Element /subelement	Repeatable	Comments	Status
Place where event took place			
<eventPlace>	Yes	https://lido-schema.org/schema/latest/lido.html#eventPlace Wrapper for information about place of event https://lido-schema.org/schema/latest/lido.html#namePlaceSet	Mandatory if applicable
<eventPlace><placeID>	Yes	https://lido-schema.org/schema/latest/lido.html#placeID Identifier for the place. Ideally references an authority file.	Recommended
<eventPlace > / <place> / <namePlaceSet> <appellationValue>	For name variants only	https://lido-schema.org/schema/latest/lido.html#appellationValue Name of place	Mandatory if applicable
<eventPlace > / <place> / <gml>	Yes	https://lido-schema.org/schema/latest/lido.html#gml Georeferences of location	Recommended
Information on materials and techniques relating to event			
<eventMaterialsTech>	Yes	https://lido-schema.org/schema/latest/lido.html#eventMaterialsTech Wrapper for information about material and technique relating to the event; may vary for different parts of the object. Alternatively, or additionally, the specification of material and/or technique unrelated to the event is possible in <objectIdentificationWrap> / <objectMaterialsTechWrap> / <objectMaterialsTechSet>. See https://lido-schema.org/schema/latest/lido.html#objectMaterialsTechSet	Mandatory if applicable
<eventMaterialsTech > / <materialsTech> / <termMaterialsTech type="…">	Yes	https://lido-schema.org/schema/latest/lido.html#termMaterialsTech A material or technique For possible values for the @type attribute, see http://terminology.lido-schema.org/termMaterialsTech_type Term	Mandatory if applicable

9 Theme / content

Element /subelement	Repeatable	Comments	Status
<subjectSet>	Yes	https://lido-schema.org/schema/latest/lido.html#subjectSet Wrapper for information about theme or content of an object Note that different entities (term, actor, place, event, object) have their own subelements.	Mandatory if applicable
<subjectSet> / <subject>	No	https://lido-schema.org/schema/latest/lido.html#subjectConcept Indexing of topics or content of the object. This may include general terms (for <subjectConcept>), individual terms (for <subjectActor>, <subjectPlace>, <subjectEvent>, <subjectObject>) or time specifications (<subjectDate>). https://lido-schema.org/schema/latest/lido.html#subjectConcept https://lido-schema.org/schema/latest/lido.html#subjectActor https://lido-schema.org/schema/latest/lido.html#subjectDate https://lido-schema.org/schema/latest/lido.html#subjectEvent https://lido-schema.org/schema/latest/lido.html#subjectPlace https://lido-schema.org/schema/latest/lido.html#subjectObject Term (for all subelements except <subjectDate>)	Mandatory if applicable

Administrative metadata

10 Rights relating to object

Element /subelement	Repeatable	Comments	Status
<rightsWorkSet>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsWorkSet Wrapper for information about rights relating to the object / work.	Mandatory if applicable
<rightsWorkSet> / <rightsType>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsType Rights or licence statement Term	Mandatory if applicable
<rightsWorkSet > / <rightsHolder> / <legalBodyName>/ <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#rightsHolder https://lido-schema.org/schema/latest/lido.html#appellationValue Name of rightsholder.	Mandatory if applicable

11 Dataset

Element /subelement	Repeatable	Comments	Status
<recordID>	Yes	https://lido-schema.org/schema/latest/lido.html#recordID Unique identification number in data supplier's (local) system.	Mandatory
<recordType>	No	https://lido-schema.org/schema/latest/lido.html#recordType Type of dataset: states whether the dataset describes a single object, collection, series, group of objects, or similar. For possible values, see http://terminology.lido-schema.org/recordType Term	Mandatory
<recordSource>	Yes	https://lido-schema.org/schema/latest/lido.html#recordSource Wrapper for information about source of dataset, usually the institution supplying the data	Mandatory
<recordSource> / <legalBodyID>	Yes	https://lido-schema.org/schema/latest/lido.html#legalBodyID An identifier for source of dataset, e.g. ISIL	
<recordSource> / <legalBodyName>/ <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#legalBodyName https://lido-schema.org/schema/latest/lido.html#appellationValue Name of dataset source.	Mandatory
<recordRights>	Yes	https://lido-schema.org/schema/latest/lido.html#recordRights Wrapper for information about rights relating to dataset.	Mandatory if applicable
<recordRights> / <rightsType>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsType Rights or licence statement Term	Mandatory if applicable
<recordRights> / <rightsHolder> / <legalBodyName>/ <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#rightsHolder https://lido-schema.org/schema/latest/lido.html#legalBodyName https://lido-schema.org/schema/latest/lido.html#appellationValue Name of rightsholder.	Mandatory if applicable
<recordInfoSet> / </recordInfoID>	Yes	https://lido-schema.org/schema/latest/lido.html#recordInfoID An identifier for the published dataset, preferably in the form of a dereferenceable URI;	Mandatory if applicable
<recordInfoSet> / <recordInfoLink>	Yes	https://lido-schema.org/schema/latest/lido.html#recordInfoLink Link to metadata record or to object presentation.	Mandatory if applicable

Element /subelement	Repeatable	Comments	Status
<collection>	Yes	https://lido-schema.org/schema/latest/lido.html#collection Wrapper for information about the (analogue or digital) collection / object selection the object is assigned to. In particular, the relation to a digitisation project can be recorded here. For examples of how to assign the subelements <displayObject>, <objectWebResource>, <objectType>, <objectName> and <objectNote>, see https://doi.org/10.11588/data/CHEPS6/PPPEMA Should an ID exist for the collection / object selection (e.g. DFG project number), the subelement <objectID> may be used, see https://lido-schema.org/schema/latest/lido.html#objectID	Recommended

12 Representations

Element /subelement	Repeatable	Comments	Status
<resourceSet>	Yes	https://lido-schema.org/schema/latest/lido.html#resourceSet Wrapper for information about (digital) representations of the object.	Mandatory if applicable
<resourceSet> / <resourceRepresentation >	Yes	https://lido-schema.org/schema/latest/lido.html#resourceRepresentation Link(s) to digital representation(s) of the object, if applicable in different resolutions and with technical information on playback (audio, video)	Mandatory if applicable
<resourceSet> / <rightsResource>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsResource Wrapper for information about rights relating to digital representation.	Mandatory if applicable
<resourceSet> / <rightsResource> / <rightsType>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsType Rights or licence statement Term	Mandatory if applicable
<resourceSet> / <rightsResource> / <rightsHolder>	Yes	https://lido-schema.org/schema/latest/lido.html#rightsHolder Information on rightsholder	Mandatory if applicable
<resourceSet> / <rightsResource> / <rightsHolder> / <legalBodyID>	Yes	https://lido-schema.org/schema/latest/lido.html#legalBodyID An identification for the rightsholder, e.g. ISIL or GND ID	Mandatory if applicable
<resourceSet> / <rightsResource> / <rightsHolder> / <legalBodyName>/ <appellationValue>	For language variants only	https://lido-schema.org/schema/latest/lido.html#appellationValue Name of rightsholder	Mandatory if applicable

2. Examples of LIDO datasets

LIDO sample datasets are available at:

(1) https://lido-schema.org/examples
(Please consult the examples for the latest LIDO version)

(2) Knaus, Gudrun; Kailus, Angela; Stein, Regine, 2022, "LIDO-Handbuch für die Erfassung und Publikation von Metadaten zu kulturellen Objekten - Band 2: Malerei und Skulptur [Anwendungsbeispiele]", https://doi.org/10.11588/data/CHEPS6, heiDATA, V1 (check for newer versions if applicable)