Manuscriptorium v. 1.0

Manuscriptorium Technical compatible

Technical compatibility of metadata

Draft version 1.2

AiP Beroun s.r.o:

Ing Karel Kučera
Ing František Šibrava
Bc. Martin Majer

Table of contents

About this document

Purpose

This document contains the basic guidelines and the data and metadata parameters for the incorporation of catalogue records (metadata) and digital copies of documents (data) in the Manuscriptorium system.

Anticipated readership

This document is intended for anyone involved in the preparation of catalogue records and digital copies of documents for incorporation in the Manuscriptorium system.

References

Reference is made in the document to the following sources:

[1] Manuscriptorium v.2.0 - analysis of the system, December 2004
[2] Manuscriptorium v. 2.0 – the complex digital document, October 2005
[3] Memoriae mundi series Bohemica, see URL: http://digit.nkp.cz
[4] Manuscript Access through Standards for Electronic Records (MASTER), see URL: http://xml.coverpages.org/master.html
[5] Reference Manual for the MASTER Document Type Definition, see URL: http://www.tei-c.org.uk/Master/Reference/oldindex.html
[6] MEdit, see URL: http://www.memoria.cz/download/medit_cz.asp
[7] TorXmlValid, see URL: http://www.memoria.cz/site_cz/download.asp
[8] jEdit, see URL: http://www.jedit.org/
[9] Emacs, see URL: http://www.gnu.org/software/emacs/emacs.html
[10] NoteTabLight, see URL: http://www.webmasterfree.com/notetablight.html
[11] Manuscriptoium – basics and compatibility: http://www.memoria.cz/docs/manuscriptorium_basics_and_compatibility_ENG.pdf
[12] Document selection and preparation of descriptions: http://www.memoria.cz/docs/manuscriptorium_document_description_ENG.pdf
[13] Manuscriptorium Image Quality: http://www.memoria.cz/docs/manuscriptorium_image_quality_ENG.pdf
[14] Manuscriptorium Technical compatible: http://www.memoria.cz/docs/manuscriptorium_compatibility_technical_ENG.pdf

Projects:

- national, see http://digit.nkp.cz/projekty/ProjektyVaV.htm
- international, see http://digit.nkp.cz/Projects/index_cz.htm

Introduction

The Manuscriptorium system is designed as a digital library of manuscripts, old printed books and other scarce documents. Like any other library, it contains a catalogue – in this case the OCHBR (Open Catalogue of Historical Book Resources) and its own digital documents which are held in the repository. The OCHBR brings together information (metadata) about physical documents (manuscripts etc.) in the form of catalogue records in XML format. The repository contains digital copies of a subset of these catalogued documents, known as complex digital documents (CDD) In principle, the Manuscriptorium system stores the metadata centrally in the OCHBR and provides access to digitised documents held in the operator's data storage facilities and in the remote storage facilities of other contributors.

Technical compatibility of metadata

So that individual contributors can contribute to the Manuscriptorium system both catalogue records and the digital data to which they relate, it is necessary to establish the fundamental technical criteria and parameters applying to the data to be entered. A contributor must achieve compatibility with the Manuscriptorium system for the contributed data and metadata. The system director will then guarantee its import to the Manuscriptorium database.
Contributors' compatibility with the Manuscriptorium system must be ensured on two levels. Contributors must ensure compatibility of the metadata they contribute to the Manuscriptorium database and where they co-operate with the system in the sphere of digital data they must also ensure compatibility of format, observing the naming conventions for filenames etc.

Compatibility of metadata format

Data for catalogue records must be delivered in XML format and UNICODE UTF-8 encoding. The XML file structure is prescribed by the MASTER standard. The generation of digital documents facilitating access to distributed data held in the contributors' data storage systems will be performed in the XML MASTER + format, described below. Image data must be stored in formats directly supported by internet browsers. These are the JPEG, GIF and PNG formats.

Compatibility of content

For it to be technically possible to store data records from individual contributors in the central catalogue database, these records must contain information as a minimum on the level of the obligatory elements of **DTD MASTER** [5] in the element `<msDescription>`.
These are the elements `<settlement>`, `<repository>` and `<idno>`.
Because this information is in practice inadequate for cataloguing purposes, a so-called minimum record has been established in co-operation with the National Library of the Czech Republic. This provides the optimal level of information a valid data record should contain (given, of course, that the information is available).

Settlement	Names the city (or other place) where the document described is stored, not the institution - the latter is catered for under Owner. `msDescription/msIdentifier/settlement`
Repository	Names the institution in which the document described is stored. `msDescription/msIdentifier/repository`
Shelfmark (Classmark, Call number, Accession number)	Contains the shelfmark or other identifier of the document (e.g. in the case of archive documents the name of the archive collection in addition to the specific classmark). `msDescription/msIdentifier/idno`
Main title	The title of the document - in the case of a collection of texts within a single document it is appropriate to give a collective title, e.g. Textus varii, Collection of Laws etc. `msDescription/msHeading/title`
Author	The author of the document or of certain of its parts. Author means the intellectual originator of the text, not a scribe, for example. It may include more than one name. `msDescription/msHeading/author`
Year of publication	Indicates the date of origin of the document. May be given as a precise date or as any time span. `msDescription/msHeading/origDate`
Language of the original document	The language in which the document was written. More than one language may be mentioned. `msDescription/msHeading/textLang`
Note	Any other data the author of the document description considers appropriate. `msDescription/msHeading/note`
Intellectual Content	Enables a brief, summarised description of the content of the document to be given (e.g. A collection of legal texts relating to the field of South German municipal law). `msDescription/msContents/overview>/p`

Illumination	Enables any information to be given about the decoration of a manuscript; appropriate also for engravings in printed books. Either a brief descriptive summary or a description referring to individual folia of a manuscript. `msDescription/physDesc/decoration/decoNote/p`
Notation	A field for data on any musical notation the document may contain. `msDescription/physDesc/musicNotation/p`
Binding	Contains a description of the binding of the document. `msDescription/physDesc/bindingDesc/binding/p`
Material	Describes the material on which the document is written (usually paper, parchment or a combination of both) `msDescription/physDesc/support/p`
Extent	Records the number of pages (or folia) in the document, including any prefaces and end-papers. It is also appropriate to mention any errors in the listing of folia (or pagination) – missing pages or duplicated folio numbering. `msDescription/physDesc/extent`
Dimensions	The dimensions of individual leaves may be given (if they are identical or similar) or any range of maximum/minimum values. `msDescription/physDesc/extent`
Bibliography	Publications relating to the document described may be noted (editions, catalogues, monographs devoted to specific works, journal articles etc.) `msDescription/additional/listBibl/bibl`

Extension of the format for the description of old printed books

Place of printing

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>printer</resp>
      <name type="place" role="printer">Place of
       printing</name>
    </respStmt>
  </msHeading>
</msDescription>
```

Name of printer

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>printer</resp>
      <name type="person" role="printer">Name of printer
       </ name>
    </respStmt>
  </msHeading>
</msDescription>
```

Place of publication

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>publisher</resp>
      <name type="place" role="printer">Place of publication
       </ name>
    </respStmt>
  </msHeading>
```

Name of publisher

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>publisher</resp>
            <name type="person" role="printer">Name of
       publisher
       </ name>
    </respStmt>
  </msHeading>
</msDescription>
```

Formal compatibility of image data

A separate document is devoted to the compatibility of image data **[13]**.

The Manuscriptorium system anticipates that digital images are made accessible to users on the internet at one or more quality levels. Normally, these are:

NORMAL the highest quality of images accessible on the internet. (at this quality level, digital images from the digitisation studio of the Czech National Library have a resolution of approx. 220 dpi and are saved in JPEG format). This image quality is essential for inclusion in the Manuscriptorium system.

GALLERY small images for the purpose of creating galleries of the individual pages of a digitised document. Manuscriptorium works with images in JPEG format with a height of 100 pixels (retaining the aspect ratio of the pages). If this image quality standard is not met, a gallery will not be created in the Manuscriptorium system.

PREVIEW preview images for navigation in the full-size image (NORMAL quality). Manuscriptorium works with images in JPEG format with a width of 200 pixels (retaining the aspect ratio of the pages). If this image quality is not met, preview images are created automatically by the system.

The system may also make images available at any other image quality level, if there are references to them in the descriptive metadata (MASTER+).

Basic conventions

To enable the effective generation of XML documents for the description of a digital copy, the basic guidelines for the naming of image files and their location in the directory structures must be observed.
Directory names and file names may contain only characters listed under the ISO646 standard

a) upper case letters without diacritics ‘A’ .. 'Z' (0x41..0x5A)
b) numerals '0'.. '9' (0x31..0x39)
c) underscore '_' (0x5F)

This is to ensure permanent transferability of data between different operating systems.
Name length is acceptable for OS Windows, Linux, MacOS and ISO 9660.

One file with an image at the given quality level corresponds to one scanned unit of the original document, identified by its classmark (e.g. for a manuscript this is typically one page). All such files are located in one directory, which may have a sub-directory structure. For access to a digital document on the internet via the http protocol this directory is referred to as the so-called “base address of the digital document” (URL).

Image files corresponding to the same scanned unit at different quality levels must be distinguished in the file name or they must be located in different sub-directories.

Image files are identified in the file name by the folio numbers or the page numbers of the original scanned unit (pages of the document).

Image file names

Two formats are permissible for file names in the Manuscriptorium system, indicating or not indicating image quality. If more than one image quality level exist for a given scanned unit the image quality must be indicated in the file names or the separate quality levels must be saved in separate sub-directories. An image file name without indication of quality has the structure

```
ppppppFFFFF.XXX
```

, and in a file name including an indication of image quality two characters specifying the quality level are inserted between the prefix and the folio/page number

```
ppppQQFFFFF.XXX
```

where

- **ppp…** s the prefix of a file name with a length of 0..20 characters identifying an original document
- **QQ** indicates image quality by means of two characters. The Manuscriptorium system adopts e.g. G0 for gallery, NO or N1 for “NORMAL“ quality level and PO for “PREVIEW“ etc.
- **FFFFF** (5 characters) – folio or page numbers, filling any empty spaces with zero characters (‘0’).
- **.XXX** is the suffix of the file name (full stop (period) + 3 characters). The suffixes JPG, GIF and PNG are permitted.

An image file name is created according to the above-mentioned guidelines; the five characters identifying a page (F) are generated according to these guidelines:

Part of the manuscript/printed book, in Czech	Part of the manuscript/printed book, in English	File name from scanner	FFFFF
Běžný list	Standard Sheet	0001r.JPG 0001v.JPG 00001.JPG	0001R folio no. 0001V folio no. 0001P pagination
Vložený list	Enclosed Sheet	ESnnn.JPG	ES001
Zpevňovací proužek	Reinforcing Strip	RSnnn.JPG	RS001
Hřbet	Spine	SP.JPG	000SP
Horní ořízka	Head Edge	HE.JPG	000HE
Boční ořízka	Side Edge	SE.JPG	000SE
Dolní ořízka	Bottom Edge	BE.JPG	000BE
Přední desky	Front Cover	FC.JPG	000FC
Přední přídeští	Front end-sheet	FS.JPG	000FS
Zadní desky	Back Cover	BC.JPG	000BC
Zadní přídeští	Back End sheet	BS.JPG	000BS

Římské číslov. přední	Front roman page	Frrrr.JPG	F001R folio no. F001P pagination
Římské číslov. zadní	Back roman page	Brrrr.JPG	B001V folio no. B001P pagination

The file name of an image in JPEG format at the highest quality level available (which the contributor will indicate, e.g. „EX“) for a page with folio no. „1r“ of a manuscript, identified e.g. by the prefix „RUK1“ may take the following form:

```
RUK1_EX0001R.JPG
```

where

- **RUK1_** RUK1_ is the prefix identifying the manuscript to which the image belongs
- **EX** is the indication of image quality
- **0001R** 0001R is the identification of the document by folio no.
- **.JPG** is the suffix of an image file in JPG format

If image file names do not contain image quality information, files with different image quality levels must be located in different directories (the directory names then carry information on the image quality of the files they contain). Thus the full path name to the root directory for a digitised physical document takes the form:

```
\QQ\ppppFFFFF.XXX
```

The file name, including sub-directories for quality levels, will in this case take the form:

```
\EX\RUK1_0001R.JPG
```

The file name does not now contain image quality information, but the file is located directly in the EX directory, which contains all image files of the given digital document at this quality level. The name then consists, as in the previous case, of a prefix identifying the digital document (RUK1_), the identifier of the digitised page (0001R) and a suffix indicating the image format (.jpeg)

If the file names contain image quality information, the files for all image quality levels may be located in one common directory for the whole digitised document.

Access to image data on the internet

To enable sharing on the internet of digital data (image files) from the web servers of individual contributors via the Manuscriptorium system, the following conventions must be observed.

Direct access

Data from contributors must be accessible via the http protocol by entering a unique address (URL) which for the above example may take the following form, e.g.

```
http://www.memoria.cz/images/RUK1/EX/RUK1_0001R.JPG
```

Similarly, if the file names contain image quality information it is not necessary, especially for small documents, for these files to be located in different directories and the address may take the following form:

```
http://www.memoria.cz/images/RUK1_EX0001R.JPG
```

Indirect access

In addition to this direct approach, it is also possible to access images via a web server extension (cgi, asp, php etc.), The URL will then take the following form, e.g.

```
http://www.memoria.cz/system/img.cgi?DocId=RUK1_&IQ=EX&page=0001R
```

where the parameters of the **img.cgi** script are:

DocId	identifier of the digital document
IQ	image quality
Page	page identifier

The script may, if required, include further parameters, e.g. authentication data in cases where the images are not freely accessible etc.

Connectors

Because the range of data storage solutions is in practice wider than the possibilities indicated by the direct approach and especially by the indirect approach, compatibility may be achieved by the building of an individual connector enabling communication between Manuscriptorium and a partner's storage facility. This connector may be attached to the Manuscriptorium server or to the contributor's server.

A connector attached to the Manuscriptorium server may be used by multiple providers.

The use of connectors is always the subject of a bi-lateral agreement between the data contributor and the director of Manuscriptorium.

Management of data storage facilities

The management of each data storage facility holding digital copies or complex digital documents is under the full control of the director of the data storage facility. The owner of the digital data bears full responsibility for the correct structure and content of the data held in the data storage facility. The manager of the data storage facility is the member of staff

responsible for the location of digital copies of a document in the data storage facility and for its correct structure and accessibility.

Compatibility control procedures

To enable the secure incorporation of contributors' data and metadata in the functioning Manuscriptorium system it is vital to ensure that the data and metadata supplied are submitted to uniform compatibility control procedures before they are incorporated in the system.

The compatibility control procedures are as follows:

The contributor provides the director of Manuscriptorium with a representative sample of catalogue data records and data prepared according to the guidelines set out in the present document. The director subjects the data supplied to an analysis of its form and content. Any identified inadequacies are reported to the contributor, who amends the data and provides the director with a new sample. This procedure may be repeated. On its positive outcome, the contributor's metadata will be acceptable to the Manuscriptorium system and can be imported.

On reaching this stage, the contributor receives a "Manuscriptorium Compatible T" certificate confirming the technical compatibility of the data and its usability in Manuscriptorium.

This compatibility is an essential pre-requisite for the acceptance of an application for a licence permitting unrestricted use of the entire contents of Manuscriptorium free of charge.