# Manuscriptorium v. 1.0
# Document selection and preparation of descriptions

Draft version 1.2

National library ČR
PhDr. Zdeněk Uhlíř

23.2.2006

## Table of Contents

# About this document

## *Purpose*

This document contains the basic guidelines for the selection of documents for digitisation, introductory information about standards governing the preparation of descriptive catalogue records and the data and metadata parameters for the incorporation of 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.

## *Terminology and conventions*

The terminology and conventions used in this document are described and defined in document [1], chapter 6.2. Conditions regarding access to data and metadata are described in document [2].

## *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, available at URL: http://digit.nkp.cz
[4] Manuscript Access through Standards for Electronic Records (MASTER), available at URL: http://xml.coverpages.org/master.html
[5] Reference Manual for the MASTER Document Type Definition, available at URL: http://www.tei-c.org.uk/Master/Reference/oldindex.html
[6] MEdit, available at URL: http://www.memoria.cz/download/medit_cz.asp
[7] TorXmlValid, available at URL: http://www.memoria.cz/site_cz/download.asp
[8] jEdit, available at URL: http://www.jedit.org/
[9] Emacs, available at URL: http://www.gnu.org/software/emacs/emacs.html
[10] NoteTabLight, available at 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

# Document selection in Manuscriptorium

Manuscriptorium comprises not only an open/union catalogue of historical (book and library) resources, i.e. manuscripts and printed books and some maps up to 1800 inclusive, but also a digital library providing access to full digital copies of this type of document. Selection of content for Manuscriptorium is therefore unlimited in terms of volume. In practice, however, limitations are imposed by financial constraints of contributors to Manuscriptorium. It is important for the selection of documents for digitisation to be consistent, i.e. the resources of the contributor should be allocated on some transparent basis. The selection must therefore be based on the priorities of the respective contributors.

There are three general principles governing the selection:

1) Documents which are characteristic or typical examples of contributors' holdings, i.e. a collection offering a profile of their historical resources, to be digitised and subsequently made accessible as systematically as possible. In this case consideration may be given to the attachment of secondary documents, i.e. commentaries, covering articles etc., though this is not a requirement.
2) Documents which are considered to be historically the most notable of the contributors' holdings, or their finest in artistic terms etc. In this case consideration may also be given to the attachment of explanatory texts, though this is not a requirement.
3) Documents which are of the greatest interest to users, whether for purposes of their reproduction in printed materials or for purposes of specific research interest (e.g. preparation of editions). In this case consideration may be given to the linking of a digital image of the original document to its full text edition, though this is not a requirement.

Where contributors have sufficient financial resources at their disposal, all the above options can be taken up. Where financial resources are more limited, however, it is appropriate to select one option only. Ideally, systematic digitisation of documents typical or characteristic of contributors' holdings is preferred (general option 1), with the addition of the most notable documents (general option 2) and those in greatest demand (option 3), but this is not a requirement.

# Document descriptions

In principle, document description precedes digitisation as such. Descriptions of digitised documents must be sufficiently detailed, as they are intended not only for cataloguing but also for linking the digital images to an electronic document, typically taking the form of a virtual book. However, the description of digital images has proved inadequate, because of the absence of a good deal of information for which a digital image (a copy) is no substitute and which only the original document itself contains. The same is true of the traditional codicological cataloguing method because it is not primarily a substitute representing the original in words, such as is undoubtedly justified in the sphere of printed works. If a representation of the original is available in electronic form as an image and the gradual addition of further information is anticipated, a presentation in words (even an interpretation) of the original is redundant from the outset, unless its purpose is to provide significant information leading to the location of the document and the information it contains. One of the advantages of preparing the description before digitisation takes place is that a simultaneous investigation is undertaken to ensure that the manuscript is suitable for digitisation and that no risks to it are involved.

The DOBM (Digitization of Old Books, Manuscripts, and Other Documents) format for closely structured description based on SGML, which was introduced in the early days of digitisation, has the advantage of being straightforward, simplifying both routine and industrial tasks. However, it has the disadvantage that it works with closely structured data on the one hand and with entirely unstructured data on the other hand; bibliographical data and data on certain readily categorisable external characters take the form of closely structured data, whereas the remainder of the data, including data on the intellectual content of the original document, take the form of free text. When the description is detailed (and therefore extensive) both basic orientation in the displayed record and more sophisticated searching become more onerous. The strict record structure has proved to be merely transitional to a form of record that will also make use of semi-structured data.

For this reason the principle of strictly structured descriptions was abandoned in a subsequent development phase, to be replaced by the MASTER standard (Manuscript Access through STandard for Electronic Records) created under the auspices of a European Project in which the National Library of the Czech Republic participated as a full partner.

The MASTER standard (initially based on SGML, subsequently on XML) enables the creation and exploitation of data, in particular semi-structured data; thus it is more adaptable, both to variations in the material to be described and to navigation and search procedures relating to displayed data. It is based on deep structuring of the content elements and on relatively free exploitation of functional elements which may relate to various horizontal and vertical locations in the structure of the full description. Only the rules of syntax are fixed. In the MASTER standard, therefore, it is possible to establish both very simple records with minimal information content and records exploring the original document in depth. This means that it has very wide ranging and flexible practical applications, adaptable to various purposes and various levels of knowledge of the material, without restricting its use in the information system.

# MASTER

The introduction of descriptions in XML format within the framework of the **MASTER** [4] project led to the initial description of documents in a free structure better adapted to users' needs, using only XML. These resources are now freely available and can be used by anyone intending to offer documents for Manuscriptorium. It is worthy of note that originally the vast majority of document descriptions were prepared by a rather extensive group of contributors, whereas now a substantial number of documents can be described by their managers themselves. Successful co-operation requires only a very brief, straightforward period of training in the formal rules which must be observed.

A record in the MASTER standard is based on a fundamental division into 1) identification of the document, 2) title, 3) content, 4) physical description, 5) history of the document and 6) supplementary information. This subdivision is further structured to enable the creation of both quite elementary records, i.e. with minimal information content requirements, and records that are rich in information content and complex in structure. Both strictly structured and semi-structured data are used, so that effective search procedures can be readily created over this basic setup. A great advantage of the MASTER standard is that it permits the integration of data from a variety of sources, thus forming a basis for a digital library on a European scale.

### *MASTER+*

MASTER+ is an extension of the MASTER standard for the creation of virtual books, i.e. the linking of a descriptive record with a sequence of digital images - copies of the pages of an original document. To create a template for a document with which digital images will be associated, it is necessary:

1) to determine how many numbered sheets/pages it contains and if this number varies from the number of physical sheets/pages (in cases where some original sheets/pages are missing, or in cases where some pages/sheets are not numbered) to identify the number of the last sheet/page;
2) to identify the extra sheets/pages (in cases where some pages are not numbered, adding them to the numerical sequence following the last preceding numbered sheet/page by means of special notation);
3) to identify missing sheets.

This template can be created by means of the special tool MEdit, which generates the correct pagination for linking to the digital images, applying appropriate notation. The more advanced MTool enables the creation of paths from the generated pagination direct to the digital images. By this means links can also be created to digital images not physically located together with the relevant descriptive records.

The MASTER+ standard, using MTool, enables straightforward integration of sources, making it possible to create a digital library in a Europe-wide framework.

# Practice in record authoring

Basic records (with minimal information requirements) can be created following a brief training period. The creation of records containing more detailed information demands some practice. The preparation of records with a high level of information content requires substantial experience, including detailed knowledge of mark-up language and its syntax. However, all three basic levels of records can be used in a single information system. All three basic levels of records can be used for searching, though to varying degrees.

A big advantage of working with MASTER and MASTER + standards is the ease of integration of insufficiently formally standardised heterogeneous data. A further advantage is the facility for flexible management in the preparation of descriptive records, i.e. the output of the authors of descriptive data can be readily varied according to current needs.

# XML editors

Descriptive records can be prepared with the aid of various tools:

1) MEdit [6] is an editor which enables the creation of records with minimal information requirements only, and mark-up options are limited. At the same time, however, it permits the generation of a basic document template (pagination). The completed records do not require validation.
2) NoteTabLight [10] enables the creation of more complex records, containing more detailed information and permitting a flexible approach to mark-up. However, it has the disadvantage of requiring validation of the records created, using the special TorXmlValid

tool. A further disadvantage is that it does not permit the generation of a basic document template (pagination).

3) Emacs **[9]** and jEdit **[8]** are sophisticated XML editors enabling the creation of the most complex descriptive records and their validation. These are tools for experienced professional users only. They have the disadvantage of not permitting the generation of a basic document template (pagination).
4) MTool [http://www.memoria.cz/mtool/] is a complex editor and while it enables the creation of records with minimal information requirements only, it does also permit both the generation of a basic document template (pagination) and the creation of a path from the generated pagination directly to the digital images. The descriptive record does not require validation; the document is checked in its entirety.

# Conclusion

The range of available tools therefore caters for users with varying levels of expertise - for those with limited experience as well as for experienced professional authors. A system of best practice has been devised for the development of an integrated digital library. Technically, as far as volume is concerned, the way has been opened to the establishment of a European digital library of historical resources.