ECP 2006 DILI 510049

ENRICH

Definition of basic conditions for sharing of large data sets in the frame of Manuscriptorium

Deliverable number	*D 5.1*
Dissemination level	*Public*
Delivery date	*30/09/2008*
Status	*Draft*
Author(s)	*Tomas Psohlavec, AIP*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
1.0	02/09/2008	Initial version.	TP, AIP
1.1	6/10/2008	Final version	TP, AIP
1.1	21/10/2008	Language correction	GL, CCP

Document Review

Reviewer	Institution	Date and result of the review
Zdenek Uhlir	NKP	20/09/2008 Sent back with minor comments
Zdenek Uhlir	NKP	24/10/2008 Document approved for submission to EC

Approved By (signature)	Date
	24/10/2008

Accepted by at European Commission (signature)	Date

1 Executive Summary

Deliverable 5.1 defines basic conditions for sharing large data sets in the frame of Manuscriptorium. The definitions originated as a summary of outcomes of preceding WP 2 and WP 5 activities and will serve not only for internal documentation of the ENRICH project principles yet also it will be possible to make the content publicly available within the ENRICH website. This will be important for associated partners interested in participation in co-building the virtual research environment by contributing with their own digital content.

CONTENT

2 Introduction

The objective of the ENRICH project is to create a base for the European digital library of cultural heritage (manuscript, incunabula, early printed books, archival papers etc) by integration of existing but scattered digital content within the Manuscriptorium digital library.

In order to achieve such a goal the project developed conditions that enable the partners (both the actual and the ones who will join us later) to bring together appropriate mass of digital content. These conditions are open to approaches that may be applied by various institutions active in the field of digitisation of rare materials (national, university and other libraries and institutions holding historical funds).

Thus open and based on simple principles the conditions impose certain requirements on the approaching partners. These requirements are summarised and explained below.

Because our experiences show that only a simple statement of requirements could discourage those incoming partners who are less developed and/or prepared for such cooperation, we provide also detailed information on the possible methods of cooperation. Each partner interested in cooperation may easily check the list of requirements in order to see if they comply with the cooperation conditions and also what will be the most suitable way for them The information generalised here is based on the information gathered from the WP2 deliverables.

At the time of creation of this deliverable all the principles described have been already applied and verified in the tasks of WP 5.

3 Sharing digital documents within the Manuscriptorium platform

3.1 Data and metadata of a digital document

First of all it's important to realize that not only the data (e.g. image files) have to be properly prepared but also appropriate metadata content must be available in order to enable sharing of large data sets.

Under the term of **data** we understand any digital representation of the original content of the document, e.g. images files with captured pages, audio files reproducing the original textual or musical content, video files etc.

Under the term of **metadata** we understand any additional information created about the original document and/or the produced digital representations. There are various metadata types: descriptive metadata, structural metadata, preservation metadata, administrative metadata etc. All these metadata types are very important for a specific field of usage and generally the better is the content of the metadata the more usable is the produced data and the higher quality digital document is produced.

As the WP5 of the ENRICH project focuses on the aggregation tasks for presentation purposes the two important metadata types are the

- **descriptive metadata**: most often bibliographical descriptions
- **structural metadata**: information on the logical structure of the original document and its digital copy, list of appropriate data files and information on their location

These two types of metadata are required to enable the sharing of data. The information either must exist or may be created during the cooperation (methods and tools are available in the Manuscriptorium system).

3.1.1 Structural metadata

The **structural metadata** are crucial for the completation of digital document. The information indicates the structure of the original document and how this structure corresponds with existing data files. The structural metadata is processed during importing to Manuscriptorium in order to know which data files belong to which document, which part of the document, in what order etc. This structural information is used when displaying a digital copy of the document in the Manuscriptorium end-user interface.

The simplest form of structural metadata of a document may have a form of a **list of images belonging to that document.** More advanced form (highly appreciated by the end-users) of structural metadata **may** include also

- information about applicable foliation/pagination which complies with the applicable foliation/pagination of the original (enables efficient cross-referencing),
- information about chapters with reference to particular folios/pages,

- any other logical structure information.

Without at least the minimal list of images a representation within Manuscriptorium is not possible. The higher quality structural metadata a document contains, the better are the presentation results.

3.1.2 Descriptive metadata

The **descriptive metadata** obviously inform the user about various properties of the original document (identification information, intellectual content, physical content, history etc) or its digital representation (availability of a digital copy etc).

Also it plays another important role: there are hundred thousands of aggregated records in the platform and this mass is searchable on the basis of descriptive metadata content - which is indexed into various search fields. Therefore the bibliographic description provides a way how to reach the document. The better descriptive metadata the more visible will be the documents. Thus the minimal required extend of descriptions (as necessary minimum required by Manuscriptorium due technical reasons) represents the information identifying the original physical document (combination of country, settlement, repository and shelfmark of the original document or similar set of information describing the location of the document).

3.2 Import of documents into the Manuscriptorium platform

During import to Manuscriptorium platform only **metadata are processed** while **data remain in the management of the cooperating partner**. The metadata are processed in order to enable searching and also composing of the digital document when it is requested by the user. Data (e.g. images) are transferred directly into the end-users internet browser and the data are never imported directly into the Manusriptorium platform.

This solution has both technical (distributed responsibility for data accessibility, mutual independence, and low cost solution) and strategic reasons (partner is not forced to give the valuable data to a third party).

This is a fundamental requirement implying the fact that all **the data files have to be reachable via the HTTP protocol** in order to be accessible by the end-users internet browser.

3.3 Formats supported for import

3.3.1 Data

Any formats supported directly by contemporary internet browsers (e.g. without requiring additional plug-in; such as JPEG, GIF, PNG for image files).

3.3.2 Metadata

Any common or uncommon structured formats: TEI P5, MARC 21, UNIMARC (and MARC-like variations), MODS, ISTC, DC, MAG and more, possibly including the METS wrapper. Also any other more or less proprietary formats.

The different formats are converted using the Manuscriptorium input interfaces into the TEI P5 format in order to enable further internal processing and presentation.

4 Summary of conditions

Based on the information above it is possible to summarize the conditions as follows:

- Cooperating partner have to be able to manage the data on its own web server (or within digital library) and make them accessible via HTTP either directly or indirectly via a script serving the data.

  Examples:

  - http://www2.bne.es:81/Enrich/ENRICH/89321/89321_CatEspana_1124.jpg
  - http://teca.bncf.firenze.sbn.it/TecaFrontEnd/servlet/readImg?RisIdr=BNCF0003465034&amp;usage=3

- No particular way of data organization is required as long as the way applied is reflected by the structural metadata content.

- The data formats have to be supported directly by contemporary internet browsers.

- Cooperating partner have to be able to provide both descriptive and structural metadata for import.

- The possible ways of metadata providing are as follows:

  - harvesting of metadata via OAI-PMH interface
  - using exports of metadata transferred via FTP, HTTP, e-mail etc.
  - using direct upload of metadata of individual documents via a dedicated Manuscriptorium interface

- The information content of the metadata have to provide the information as described in the definitions above, the required minimum is:

  - appropriate list of data files (e.g. images) and their URLs (the URL may be either explicitly indicated in the list or it may be computable on the basis of available information)
  - basic original document identification

- The structural and descriptive metadata may be provided using a single format (e.g. using TEI P5, METS wrapping etc) or it is possible to perform a two-step import - processing separately the descriptive metadata (e.g. MARC records export) and the structural metadata (any structural metadata format as applied by the partner)

5 Ways of cooperation

The above conditions for cooperation are generally easy to fulfil. Depending on level of readiness of a particular partner following general ways of cooperation are possible:

- **Advanced digitisation projects operating digital libraries equipped with the OAI-PMH interface:** such partners provide Manuscriptorium with descriptive and structural metadata via one of their available OAI profiles and usually there are high quality metadata available. Metadata are harvested and processed within Manuscriptorium using individual input interface, so called Connector, which is prepared according to the metadata properties.
- **Advanced digitisation without the OAI-PMH interface:** available metadata are processed using the individually prepared Connectors for partners with larger number of documents (where Connector creation is reasonable). The difference between OAI-PMH enabled partners is in the method of transferring the metadata to the input of Manuscriptorium.
- **Starting or smaller scale digitisation projects**: where no appropriate metadata exist or where only smaller amount of documents is available (where it is not reasonable to prepare individual connector) it is possible to use Manuscriptorium dedicated tools to create and transfer the metadata content. These tools are:
  - M-Tool: an application which enables to create the descriptive and structural metadata
  - M-Can: on-line application which enables upload of the metadata into the Manuscriptorium environment, check of correctness and subsequently transfer for import
- **Large-scale digitisation projects without structural metadata**: where no structural metadata exist it may be possible to create the metadata in an automated way (using some of the commonly available generation tools) and process the results as structural metadata within dedicated individual Connector.

It is supposed one of such cooperation ways will suit to majority of possible future partners. It is possible to agree a combination of these ways or setup individual modifications where necessary.