ENRICH: Building a European Digital Library of Manuscripts

Adolf KNOLL
National Library of the Czech Republic, Prague
adolf.knoll@nkp.cz

INFORUM 2007: 13. konference o profesionálních informačních zdrojích
Praha, 22. - 24.5. 2007

Abstract:
The Manuscriptorium Digital Library provides access to catalogue data, fully digitized documents, and selected structured full texts from more than thirty various memory institutions in country and abroad. It is the largest digital library of manuscripts operated by any national library in Europe. Thanks to the new EU project ENRICH. Manuscriptorium, as a product based on 15 years of cooperation of the Czech National Library and AiP Beroun Co. Ltd., will try to aggregate similar information from many countries in Europe to provide seamless access to already dispersed historical documents. The works on which the Manuscriptorium is based were the main reason why the National Library of the Czech Republic was awarded the UNESCO Memory of the World prize in 2005 as the first institution ever. The paper gives an overview of plans and goals to be achieved during the coming two years.

When in 1992, a few people from the National Library and the company whose today's name is AiP Beroun Ltd. had decided to try to comply with a request coming from UNESCO for production of a CD concerning old books and manuscripts, none of us could know that some more years later this would lead towards consideration to aggregate similar content from more European institutions and projects under a digital library interface that got the name of Manuscriptorium. So this is our story in brief.

Manuscriptorium is the largest manuscript digital library operated by any European national library. It contains data from more than thirty institutions from the Czech Republic and abroad. These data are catalogue records, images, and recently also structured historical full texts. The volume of available digitized pages is ca. 800,000; they represent mostly manuscripts (about 65%), but also old rare printed books and historical maps. The proportion of the collections of the National Library is about 55%, while as to foreign institutions; we have mostly catalogue data (but also images) from the countries such as Slovakia, Poland, Turkey, Lithuania, Hungary, and in a very small proportion also from some other ones.

Behind the Manuscriptorium Digital Library, solid standardization issues create the platform for any further development. The digital documents comply with a complex public XML schema[1] whose descriptive part is based on the special MASTER format for electronic description of manuscripts, developed in an EU project of the same name in the beginning of the new millennium. The structural metadata reflect our long-term experience with the work with digitized manuscripts in 1990s, while the technical metadata are based mostly on the

[1] http://digit.nkp.cz/MMSB/1.1/msnkaip.xsd

NISO Z39.87-2002 standard called Data Dictionary for Still Digital Images[2] and the DIG35 Metadata Specification standard issued by the International Imaging Industry Association[3].

The image representation is mostly done by up to five images for one digitized page; these images are different as to resolution, compression quality factor, and format. In fact, the web formats are recommended, i.e. JPEG (mostly used for low and normal user quality images) and GIF or PNG (for black-and-white, thumbnail gallery images, and preview images – for the latter ones sometimes JPEG is used, too). For digital representation of historical maps the MrSID format (*.sid) is applied in combination with the Lizardtech Image Express server, which enables comfortable work with large data files of maps. The full text is structured following our own DTD[4] based on the TEI standard.

The digitization for Manuscriptorium in the Czech Republic is supported by a programme of the Ministry of Culture, which issues annual calls for proposals used by libraries of various institutions – incl. monasteries, castles, museums, or archives – to get up to 70% of funds for digitization of concrete titles from their collections.

As the manuscripts and old printed books of the same provenance are dispersed in so many collections in all over the world, we started to try to aggregate similar content also from outside of our country. We had to solve various standardization issues to be able to accept data from different resources. For the catalogue records, this meant mostly preparation of tuned conversion rules/transformations for cooperating institutions, especially from various MARC applications,

into our internal MASTER format. However, not only descriptive metadata are necessary; we would like to gather - under the same interface - also the visual representation of digitized documents to provide seamless access to dispersed resources for our users.

The manuscripts in other institutions are represented digitally in various manners: the most frequent approach is that the institutions have only the catalogue data, mostly in various clones of MARC formats, some of them also in MASTER, while other ones in some formats that are very specific. From practice, we know that the same format does not mean the same approach so that even the general transformations MARC21/UNIMARC > MASTER must be checked when having concrete results and adapted. We call such adapted transformations *connectors*, because they are able to connect the partner's metadata with our database. We think that the variety of formats is not so much a problem as it can be imperfect application of descriptive rules. In many cases, it makes sense to write a specific conversion utility when the collection is important and the application of proprietary rules is consistent.

In general, the institutions are willing to share their catalogue data, but the problems arise when we start to talk about sharing the visual representations of originals. The reasons for this may be of cultural and political origins or technical. Sometimes, it seems that it is a mixture of both.

For cultural reasons, we base the aggregation of various types of content on shared Internet storage of image or other data. In this way, we just need to upload and index the correct metadata file describing the whole manuscript with working structural links leading to

[2] Data Dictionary - Technical Metadata for Digital Still Images. Released as a Draft Standard for Trial Use June 1, 2002 – December 31, 2003. Recently published as a working draft version 1.2 (Working Draft, 1.2 March 27, 2007 - http://www.loc.gov/standards/mix/docs/NISODataDictionary_March2002.doc)

[3] http://www.i3a.org

[4] http://digit.nkp.cz/MSSFullText/DTD/1.00/mss-fulltext.dtd

existent images representing the digitized book. In such a distributed way, for example, the documents from the University Library of Bratislava, Slovakia, are made accessible through Manuscriptorium, alongside with the data from the Manuscriptorium central image bank. Thus, the central digital library database will be connected to an extended number of image banks in many institutions at home and abroad. In some cases, the libraries will enable to harvest only descriptive metadata to build a central Manuscriptorium index without provision of structural metadata making possible display of images from remote resources in the uniform Manuscriptorium interface. In this situation, the user will be navigated to respective outer resources provided their presentation interface exists. In many cases, the user will have the choice to access the same images in the source digital library or in Manuscriptorium whose added value is richer virtual context thanks to availability of many different resources.

For technical reasons, more strategies should be developed, as only a few partner libraries will be able to communicate both descriptive and structural metadata via the OAI communication protocol. In some cases, MARC records exist and are connected with images, in other cases there are no Internet presentations of digitized manuscript materials. It means that in the latter cases, necessary assistance should be provided to let the institutions come aboard the digital library. For this purpose, easy-to-use authoring, validation, and testing tools will be provided to enable certain degree of self-service for the partner institutions when sharing data with others in Manuscriptorium. This assistance is given through free downloadable tools from which the most complex one is M-TOOL, which structures the digital manuscript following the applied XML schema and enables to generate correct links to any images of manuscript pages placed on any server on the web. A *Manuscriptorium for Candidates* is providing an interface that enables the partners to upload their XML files and test the work with the digital document in an environment similar to the real Manuscriptorium. The administrator can then decide from here which documents will be uploaded into the official digital library.

However, there are even more tasks to be done:

Standardization of shared metadata

The original MASTER format for electronic description of manuscripts was developed on the P4 platform of the TEI (Text Encoding Initiative) standard. Recently, TEI announced the upgrade to P5 platform so that there is a necessity to bridge the two versions to enhance the possibilities of compatibility of native MASTER descriptions among themselves. Furthermore, it will be wise not only to transform the source descriptive metadata into the Manuscriptorium internal MASTER format, but also to enable a parallel existence of descriptions in the source formats, be it various MARC variations or others. Thus, it has been decided to enable the containerization of the metadata following the METS recommendations in order not to invent our own approaches. We call the METS-like format KDD that stands for Complex Digital Document (Komplexní digitální document). This fact does not mean the former format will be changed, it will continue to be supported as an authoring format in which the digitized documents should be structured to be acceptable for Manuscriptorium.

Manuscriptorium today is harvestable via OAI protocol, but we will implement into it also the OAI harvester to be able to offer more complex information to our users about the documents that are interesting for their research. More work will be also necessary in the field of UNICODE treatment as a condition for a deeper internationalization of the digital library.

The workpackage concerning all this work will be coordinated by the Oxford University Computing Services.

User personalization

Manuscriptorium wishes to be rather flexible when responding to users' needs; therefore, their deeper analysis will have to be done. It will lead to better bibliographic searching capabilities and deeper search opportunities operated on both metadata and full text indexes. The two main goals are:

- creation of thematic collections for users; they will reflect the needs of various user communities following such criteria as provenance or character of the documents, specific user groups (for example, a special *Manuscriptorium for Schools* has been already successfully created to reflect the needs of secondary schools in the Czech Republic), or ownership of documents;
- creation of virtual documents by researchers; the user will be offered the opportunity to compose his/her own documents from the analytical digital objects existing in Manuscriptorium beyond the entities copying the integrity of digitized physical volumes; in this way, for example, illuminations or parts of texts from similar schools and scriptoria can be grouped together for teaching purposes, while their behaviour will continue to be in line with the Manuscriptorium presentation habits.

This workpackage will be coordinated by the University of Florence Media Integration and Communication Centre.

Personalization for contributors

To make the cooperation of content partners easier, we must meet their needs; for this, we must know in which form they will be able to cooperate. From our experience, it is evident that not so many institutions are running digital libraries for such old materials and that only a very limited number of them will be able to cooperate via OAI on the level of all the necessary metadata groups. It seems that a lot of work will consist in the individual preparation of connectors for partner institutions to automate the necessary compatibility transformations from their environment into the common one. From this point of view, the connectors will serve both the possible on-line harvest of metadata and their batch processing from the uploaded cumulative output from partner resources. It is also evident that the basic support for newcomers, i.e. those who digitize without knowing what to do next, but willing to be visible in Manuscriptorium, is necessary in form of easy-to-use tools without knowledge of specific things such as XML or TEI.

This workpackage will be crucial for expanding Manuscriptorium into Europe and the world. It will be led by the Czech AiP Beroun Co. Ltd.

Multilingual and sophisticated user-friendly access

The questions of multilingualism are inherent to any European project, because it must cope with many linguistic environments. They are about Manuscriptorium interfaces in partner languages, about possible multilingual and translated search in that large variety of terms in so many different languages – even now, they are substantial portions of data not only in Latin, (Old) Czech, or German, but also for example in Turkish or Lithuanian. Also the preparation of individual linguistic interfaces and related tools should be automated as far as possible and, in plus, for better search results, more sophisticated processing of queries should be enabled with application of ontologies. As there are some partial results in the ontological area from the EU project VICODI in which several partners were involved some time ago, it would be wise to try the reuse the work already done there.

This workpackage will by coordinated by the French Systran Company.

Of course, there will be some more work in assessment of the conditions of cooperation in the starting period and also during various stages of the project incl. testing, validation, usability or adaptability of prepared applications. A lot of attention will be paid to enlargement of the cooperation onto new associated partners. In all these activities, the most active partners and workpackage leaders will be The Vilnius Institute of Mathematics and Informatics, National Library of Spain, and the coordinators: National Library of the Czech Republic and Crossczech Co., Prague.

Other important technical partners are also Nordisk Forskninginstitut at Copenhagen University, Computer Science for Humanities Centre at Cologne University, and Poznań Supercomputing Centre, Poland.

The full content partners are national libraries of Czech Republic, Spain, Iceland, and Italy (Florence) together with university libraries in Vilnius, Wroclaw, and Budapest (University of Technology and Economics), Arne Magnússon Foundation in Reykjavík, and Monasterium project of the St. Pölten Diocese in Austria.

Many content partners come also from other countries and institutions: Romania (Institute for Cultural Heritage CIMEC and Carol I University Library in Bucharest), Serbia (National Library and Institute of Mathematics and Informatics in Belgrade), Turkey (National Library), Moldova (National Library), Macedonia (Faculty of Natural Sciences and Informatics in Skopje), Sweden (Royal Library), and Kazakhstan (National Library). New additional associated partners are welcome.

Even if a lot of work is ahead, it will aggregate a greater part of digitized manuscripts in Europe, as the richest owners of digitized manuscripts are national libraries of the Czech Republic, Iceland, and Serbia, while substantial content is available also in other institutions, such as the Romanian CIMEC, National Library of Italy in Florence, or University Library in Wroclaw.

The partners will not only contribute to the common shared resource, but they will also bring the large quantity of data to their users, and they will have the opportunity to have quickly their own digital library, derived virtually from the big Manuscriptorium (http://www.manuscriptorium.com).