Преглед НЦД 5 (2004), 63–66

Adolf Knoll
(National Library of the Czech Republic)

DIGITIZED CULTURAL HERITAGE

Abstract: We present the principles of two Czech national digitization programmes: Memoria and Kramerius. The presentation concerns the question of imaging and also that of complex metadata formats developed for digitized documents. Both programmes offer rare library materials – manuscripts, old printed books, historical maps, and endangered acid-paper periodicals and monographs on Internet.

Keywords: digitization; imaging; metadata document formats; access applications

1. Digitization programmes

Today, two national digitization programmes in the area of libraries are running in the Czech Republic and there are also some other partial projects in individual institutions or other areas. It can be said that the most experience has been acquired in digitization of rare and endangered library materials, such as manuscripts, old printed books, historical maps, newspapers and journals, old monographs, etc. There are also projects that concern other materials such as legal documents (laws and regulations) in the Parliamentary Library, glass photographic stereo negatives in the Moravian Technical Museum in Brno, or the state-supported project of digitization of the Supraphon Company audio archives. There are also small partial initiatives in the museum area.

Nevertheless, only libraries have managed to have large digitization programmes that are able to support annually a considerable number of projects. These programmes have their origins in the research and development activities of the National Library of the Czech Republic that were developed in co-operation with several smaller Czech software companies. It was especially this kind of co-operation with local SME, which was the key for successful implementation of modern technologies in Czech libraries. This co-operation dates back to early 1990s when both libraries and local SME wished to play important roles in their areas. In successful cases, both parts were able to grow together: this fact enabled libraries to cope quickly with implementation of information systems, retrospective conversion, or digitization, while the companies specialized to cover these areas and work for libraries.

The most important role in digitization of library materials was played by AiP Beroun Ltd., known formerly as Albertina icome Praha Ltd. The co-operation started in 1992 with preparation of the first UNESCO Memory of the World digitization pilot project. The company and the National Library are today equal and widely recognized partners who are leading the so-called Memoria programme. During the years past, a rather good development team has been created and it can be said that the contribution of the company is much more than only the technological one, while the contribution of the library brings also important technological solutions especially in the metadata area.

Under the Memoria programme, dozens of Czech institutions – public libraries, museums, archives, and monastery and castle libraries - have digitized until today more than 1,000 manuscripts and old printed books that may represent up to 550,000 pages transformed in high-quality images. The annual production today is between 105,000 – 120,000 pages.

The second national digitization programme is called Kramerius and it has been partly built on the experience acquired in Memoria. However, Kramerius is more specific: its goal is to safeguard and preserve the acid-paper library materials, especially newspapers, journals, and brittle monographs. It started in routine in 2000, while the routine digitization in Memoria began in 1996 already. The endangered documents are filmed first and then the microfilm is scanned. The role of microfilm is to preserve the original information, while the role of the digital output is to make this information largely accessible. Kramerius is able to film ca. 1,000,000 pages of documents annually and to digitize up to 400,000. It may have ca. 1.2 or more million pages in digital form nowadays.

The original National Library digitization projects were expanded especially in 2000 thanks to the support of the Ministry of Culture and its programme Public Information Services of Libraries within the framework of which Memoria and Kramerius are sub-programmes that launch annual calls for proposals. The applicants can claim up to 70% of cost in their projects, which are evaluated by ministerial committees composed mostly from the specialists who work with historical documents and whose aim is to consider the necessity of larger availability of the titles proposed for digitization.

While the filming takes place in several institutions, the digitization for both programmes is made centrally in the National Library. The Memoria digitization work is entirely outsourced and done by AiP Beroun Ltd. that ensures smoother operation and stable quality.

2. Imaging

The production of images is primarily given by the devices that are used: two Betterlight and one Cruse scanners for Memoria and two microfilm scanners for Kramerius. The archival images are today only in JPEG format, standard encoding and very light compression ratio that is not going beyond the quality factor 12 as set up in Adobe Photoshop.

From these images, the user quality images are made as follows:

- in Kramerius the archival images are converted into DjVu (*.djvu) format for which several standard situations have been defined so that the DjVu parameters can be set up automatically in function of objects that are found in the original documents (text only, text and simple graphics, tiny details and photorealistic objects, etc.);
- in Memoria each image serves for production of a set of user images of lower quality: user quality with visible MMSB (it stays for Memoriae Mundi Series Bohemica) watermark, low quality that is still good for normal reading and study, black-and-white image in order to support reading of texts (the zones having various quality parameters have been balanced before dithering onto two

colours), preview image, and gallery (thumbnail) image; the black-and-white and gallery images are in GIF, while all the other quality levels in JPEG.

The JPEG images in Kramerius are in 256 shades of grey, while the JPEG images in Memoria are in true colour with the exception of older textual manuscripts, which were digitized with an older Kodak camera.

A specific type of digitized material in Memoria is the historical map and outside of the programme also other types of larger originals – especially plans – are digitized, too. For this type of material the user image is in the MrSID (*.sid) format, which enables fast display and navigation in large data files. The uncompressed historical maps have a typical size that goes from ca. 100 MB up to almost 0.5 GB, which is a limit for the MrSID encoder we possess.

While in Kramerius the most important feature in imaging is the correct cutting of unnecessary parts and free rotation-like correction of images taken from microfilm, in Memoria there is a lot of stress on long-term platform independence of correct rendering of colours in display and printing. For this purpose, some special techniques are used, as for example, vacuum-based fixation of digitized individual folios and pages, calibration of scanners and storage of image information via ICC profiles, calibration tables (printed with permanent colours on permanent paper, scanned with the original, and stored physically with the off-line copy), and additional imaging metadata.

3. Metadata

Since almost the very beginning of our digitization activities, we had been paying a lot of attention to descriptive metadata, but it took us certain time until we wrote a very complex metadata format for the whole digital document. It became unavoidable when we decided to digitize in routine in 1995.

The first platform on which we created the e-doc format for manuscripts and later also for digitized periodicals and even audio documents was an extended HTML language. The extension was made by completion of the HTML 2.0 DTD with several few elements that enabled description of contents. However, these elements were very general and they needed to be further specified on the basis of appropriate descriptive rules. This specification was made in a special SGML file that played a role of a concrete document type definition of the 2nd level, e.g. for manuscripts and old printed books; it contained in fact all applicable elements for all description levels (book, page, issue, volume, etc.). This definition SGML file was also controlled by a special DTD.

Such an approach gave us a lot of flexibility to implement any descriptive rules or good practices. Furthermore, it made possible basic browsing and viewing in web browsers, while the content-oriented elements could be used for indexing and more sophisticated navigation in the digital document in special access tools.

From today’s point of view we tried to solve something that was done later and with great success in XML: easy access in common tools combined with description of contents. However, the display in browsers was not done via any style formatting tools, but in a rather simpler way that consisted in duplicated mark-up of elements in the same file, where each value was marked up formally with normal prescriptive HTML tags and simultaneously also its content with the special so-called DOBM tags that were disregarded by browsers.

The further development of this e-doc platform was mostly influenced by a rather fast evolution of descriptive rules or approaches. In the manuscript area, the TEI-based MASTER format was agreed in Europe in the beginning of the new millennium, an attempt to write an e-doc format for digitized periodicals appeared in Europe, too (EU DIEPER project), and new approaches were prepared in the area of technical metadata for description of data files, especially still digital images.

It is interesting to observe that it was not so much the appearance of XML that made us rewrite the complex e-doc formats for several types of digitized documents as the necessity to adopt new standards and to place all metadata descriptions on a higher-quality level. The unsatisfactory quality was caused initially by two factors: lack of complex standards in the area and lack of skills and knowledge in content departments of the library in the moment when we launched routine digitization programmes. It took some more time until really advanced approaches were asked for implementation by the library departments; these were delayed in comparison with understanding of library problems by technicians.

As of today, the new DTD have been prepared for the types of digitized documents as follows: manuscripts and old printed books (applied also for historical maps), digitized old periodicals, digitized monographs, and museum objects. With the exception of museum objects, all the DTD have been implemented first into access systems and secondly also into authoring tools (under way now). A DTD for digitized audio documents is going to be written in the near future, while a TEI-based DTD for full texts of medieval and early modern manuscripts has been prepared, too. All our DTD are available from the http://digit.nkp.cz server.

These DTD are based on modern cataloguing practice and agreed standards; the DTD for manuscripts contains also a large entity for description of still digital images in order to ensure their true rendering over time.

4. Access

In the beginning, we had to cope with a period of off-line access and distribution of CDs, but our philosophy changed with growing importance of Internet and better connectivity of institutions and individuals.

However, it was only last year and for Kramerius even only now that we launched real access applications for Internet access. They were both developed by Czech companies. Memoria application is based primarily on our national bibliography access software (created by AiP Beroun Ltd.), while Kramerius was developed for us by Qbizm Technologies Inc. Their roles differ, because Memoria is also a shared catalogue of historical collections, while full records for Kramerius items are in OPAC.
Access to the Memoria Digital Library is always licensed, while the non-licensed users can use only the catalogue and small preview images. The copyright-free materials in Kramerius are available for anybody, but there are also many documents to which author rights apply and from this reason access to them may be restricted to certain sites only, usually document owner institutions.

The Memoria Digital Library is at the URLs:
http://www.memoria.cz or http://www.manuscriptorium.com,
The Kramerius Digital Library is at the URL:
http://kramerius.nkp.cz.