LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, printed in Germany

Digital Library for Access to Rare Materials. From Pilot Projects to National Digitisation Programmes

by ADOLF KNOLL

INTRODUCTION

More than ten years ago, a first image digitisation product was published by the National Library of the Czech Republic in co-operation with an almost unknown small software company. It was a CD-ROM requested by UNESCO for its Memory of the World programme. After that there were many tests concerning digitisation devices and another two CD-ROM projects were completed to present two interesting manuscripts with rich description metadata related to all the structural levels of the book, including full text. In 1995 – 1996, our first digitisation centre was built to process manuscripts and old printed books. Later, it was upgraded several times with additional devices such as scanners, book cradles for careful document fixation, sophisticated camera stands, microclimate control units, etc. Simultaneously, many software tools were developed to assist the digitisation production. The co-operation with the company became crucial for further development of this activity and both the library and the company grew thanks to each other.

In 1999, another project started to output digital data: it was based on the idea of safeguarding acid-paper library materials through preservation microfilming combined with digitisation of microfilm for access purposes. In order to be able to store large amounts of image data produced by this second programme, a mass storage facility was plugged into the digitisation production line. It was based on robotic library using magnetic tapes as storage medium. Successively, the digitisation programmes began to be used also by other Czech institutions holding rare collections of documents: other libraries, museums, archives, and even castle libraries and church institutions.

Today, from all the digitised documents almost half of them are from other institutions than the National Library. This is an important success that has taken place thanks to support of the Ministry of Culture through specific grant programmes. Interested institutions can ask the Ministry for up to 70% of the total digitisation cost for their projects with which they apply in calls for proposals.

PUBLIC INFORMATION SERVICES OF LIBRARIES PROGRAMME

Nowadays, the framework for support of operation of new technologies in libraries is called Library Public Information Services Programme (LPIS). It has nine sub-

programmes targeting such activities as enhancement of Internet access to information through public libraries, training of librarians, retrospective conversion of catalogues, or Memoriae Mundi Series Bohemica (started in 1996 as digitisation programme of rare library materials) and Kramerius (started in 2000 as a preservation microfilming and digitization programme of endangered acid-paper materials). Simultaneously, the National Library is engaged in several research and technology development projects and programmes, where it co-operates with many small and medium enterprises to solve crucial questions for further development of technologies.

The volume of our annual digitisation production depends on the funds that are available in the sub-programmes. In the last three years, we were able to process more than 100,000 pages of manuscripts and ca. 300,000 – 400,000 pages of acid paper materials yearly. We expect to have – in the end of 2003 – ca. 500,000 digital pages of manuscripts, old printed books, and maps and ca. 1,100,000 pages of digitised acid paper materials. The percentage of old printed books and historical maps digitised in the first programme is relatively low, as the main materials of interest are old rare manuscripts.

METADATA FRAMEWORK

When preparing the routine production of digitised manuscripts, one of the biggest concerns was the relationship between the produced documents and changing software and hardware environments. As almost all the projects of the time, first digitised images from our collections were rather strongly bound to handling and access tools that had been written for concrete platforms. However, soon it was evident that we had to find a more durable solution. In 1995 XML was not yet available, but at least there was the SGML platform. We had to find a way between the complexity of the pure SGML and the accessibility of documents encoded in the HTML. The problem was how to mark-up the contents and format it at the same time, while using for the basic work a normal web browser. The solution was an enlargement of the HTML DTD with introduction of a few content-oriented elements to enable the mark-up of contents. Thus, a kind of hybrid format was created containing the formal mark-up for display together with content-oriented elements. From the point of view of the user, the appearance was that of a labelled format in which each element was labelled by its name, while the hidden exact mark-up was used for further processing.

The general format was called DOBM and it was rather flexible so that it could control also other types of digital documents. In total, three concrete applications were written and applied in practice: manuscripts and old printed books, periodicals, and audio documents (applied only in a limited number of cases). The format was also used for creation of short catalogues of Arabic and Persian manuscripts. In order to enable the work of specialists with the DOBM format, a set of tools called ManuFreT was prepared. It enabled generation of appropriate templates for metadata entry, structuring of metadata

into correct DOBM structures as well as local sophisticated access to digitised manuscripts. The version 2 of DOBM (applied since 1997) was presented with all the definitions and tools on a CD-ROM edition summarizing all our rules and experience in this domain. The CD-ROM Digitization of Rare Library Materials: Storage and Access to Data (Knoll, 1997) was recommended by UNESCO as an appropriate approach to follow in January 1999.

There was also a lack of agreements on e-doc formats concerning digitised library materials; therefore, we limited the mandatory set of metadata only on very basic bibliographic and structural elements. It was not our intention to double the mission usually carried by library OPACs. Nevertheless, during the 4th Framework Programme, two interesting projects appeared that concerned in different ways the work with digital copies of original documents: MASTER (Manuscript Access through Standards for Electronic Records) and DIEPER (Digitised European Periodicals). We were one of the main participants in the MASTER project of which the most important output was a standard for the bibliographic description of manuscripts. After the project had been completed in June 2001, several institutions prolonged the work based on the MASTER DTD. One of them was also our library, where the MASTER format was considered appropriate for cataloguing not only manuscripts, but also other historical materials, old maps included.

Simultaneously with the works on wider introduction of MASTER approach into the Czech practice, we decided to implement the DTD also into the e-doc format for the digitised manuscripts and old printed books. The result is the complex DTD for control of digitised manuscripts and similar historical documents [1]. This new DTD defines the format into which all the metadata from our digitised manuscripts are being converted nowadays. Simultaneously, the bibliographic descriptions are being enriched so that they can be included into the database of historical documents. In fact, digitisation has raised in this domain the interest to create and build a shared catalogue of historical documents – Manuscriptorium - from which also digitised copies will be available. We have been trying to persuade also other institutions than those, which take part in the digitisation programme, to contribute with their records. It is important to observe that in spite of a possible national (union) potential of the database, the interest to co-operate is larger: we are about to reach concrete agreements with several foreign institutions among which the participation of the Wroclaw University Library (Poland) and Bratislava University Library (Slovakia) are rather sure. In this way, the catalogue may have an important regional aspect and it may really serve as a good basis for creation of a good virtual research environment in this domain.

A similar development is taking place in the area of digitised periodicals, where the old DOBM application is being replaced by a new XML structure today. Here, a new DTD was written on the basis of our analysis of the DIEPER structure, DOBM approach, good cataloguing practice, and existing needs [2]. In this case the metadata description does

not aim to replace the current OPAC data. In the new access application, both sources will be complimentary, while pointing to each other from concrete records to concrete records. The new application for access to these data is being developed nowadays and, in addition, it will serve for access to other digitised materials, such as museum objects, digitised acid-paper monographs, or digitised audio objects. The preparation of the necessary DTDs is under way; at this moment there is a draft of a DTD for a museum object [3]. This DTD is inspired by proximities of identification descriptions existent in library and museum worlds as well as by several international approaches as, for example, the RLG REACH Element Set or the French *Systèmes descriptifs des objets mobiliers* (Magnien, 1999). The latter approach has been very important, because it concerns the processing of technical objects that are of special interest for us due to a special project of safeguarding the flooded documentation of the history of Czech aviation and architecture in the National Technical Museum. The development work in this project is carried out by the National Library of the Czech Republic thanks to a special grant of the U.S. Andrew W. Mellon Foundation for technology development to assist those institutions the collections of which were damaged during the floods in August 2002.

DATA SOLUTION

Data are the basic carriers of the reformatted information: be it image or other types of files, such as text, audio, or video. It is the projection of the original into virtual space, while metadata are the help information added from outside to digitally transformed originals. We started with uncompressed bitmaps in the very first projects and with lower space and brightness resolutions, but we introduced the JPEG image into our projects very soon. Nowadays, the manuscripts and other historical materials are scanned in 16.7 million colours or 256 shades of grey in case of textual materials in which the colour does not play a decisive role, while all the microfilms are scanned in 256 shades of grey after some experimentation with the bi-level (black-and-white) image. The images are then stored in JPEG under very moderate compression ratios (up to 12 in the Adobe Photoshop) to avoid significant loss of data through compression.

We have experimented also with new emerging solutions of compression algorithms, such as clones of JBIG2 or wavelet true colour schemes as well as with mixed raster solutions. Being uncertain about the future development and expansion of these approaches, we concentrated on data delivery segment; while for archiving the classical ISO solutions are preferred (mostly JPEG). Our tests and development of various solutions and tools have made us to concentrate on very few products that we consider viable and possibly appropriate for selected delivery solutions on Internet: mixed raster content formats DjVu and LuraDocument (especially the new LuraDocument.jpm) for scanned content (mostly newspapers) and the true colour compression solutions MrSID

and JPEG 2000 (JP2). The LuraDocument and JP2 applications have been developed and are marketed by Algo Vision LuraTech GmbH, while the Dive and MrSID are marketed today by LizardTech. The mixed raster content (MRC) solutions split the image into foreground and background layers that are compressed separately with most appropriate compression schemes. The foreground layer itself is split into a 1-bit (textual) part and its colour background. The 1-bit part is compressed with JB2 scheme in DjVu (a clone of JBIG2) or with another proprietary 1-bit compressor or even with the CCITT Fax Group 4 in LuraDocument. The JB2 scheme enables also a very efficient lossy compression and it is the best 1-bit lossy compression solution on the market today. The background is compressed in DjVu in the wavelet IW44 proprietary scheme, while in LuraDocument in the wavelet LuraWave scheme. Here the LuraWave image is superior to IW44. The background in LuraDocument has one layer, while in DjVu it has four layers that present a lot of requirements for computation power during the compression, while it offers a kind of successive display of image during download.

The DjVu is the most efficient solution available in this domain, also thanks to a very good management of pre-processing compression routines, especially as to establishment of thresholds between foreground and background (Knoll, 2001) and the well-advanced solution of the so-called soft pattern matching technology in the 1-bit domain (Knoll, 2000). On the other hand, the LuraDocument solution is just shifting to a real ISO-based platform in the LuraDocument.jpm product, where the 1-bit layer is compressed losslessly in CCITT Fax Group 4 scheme (well established and widely used) and the colour information in well-mastered JP2 (JPEG 2000). As to the marketing philosophy, the German company is much more flexible than the U.S. Lizardtech, but - on the other hand - for DjVu there are also free non-Windows-based compressors available outside of Lizardtech [4]. The future will surely show better the viability of both competitors. As for us, our decision is to use DjVu as access format of digitised periodicals and similar materials in which the advantage of splitting images into text and background layers plays the decisive role. At present, a mass conversion of up to 1,000,000 images into DjVu is taking place. Both for DjVu and LuraDocument, there are freely downloadable plug-ins or ActiveX components for their handy integration into the major web browsers.

In the pure true colour area seen from the Internet access point of view, the situation is not so evident in favour of wavelet formats. It is necessary to put aside the MrSID format that is excellent for huge image files that appear in the library world especially after scanning the maps. Here, MrSID is a good choice even if it costs additional money due to unfavourably changing marketing policy of Lizadtech. Another story is the JPEG 2000. It was a long-awaited solution after many attempts to apply the wavelet compression into digital imaging and to offer a good format for it. In this domain LuraTech was always a leading company, which was achieving the best results with its older, but very good, LuraWave format [5]. It also offered the first and very good solution for the JPEG 2000 (JP2 format) that is implemented today in several widespread tools (Paint Shop Pro 8, Irfan View, Adobe Photoshop in case the plug-in is purchased

from LuraTech). We considered the application of JP2 for access to digitised manuscripts to take advantage of better compression efficiency in comparison with the classical JPEG. However, after conversion of ca. 140 manuscripts into JP2 and evaluation of efficiency and quality of results, we have recently abandoned the JP2 for this time being.

The reason for this consists in comparison of artefacts produced by the applied compression algorithms. The classical Discrete Cosine Transform (DCT) in JPEG produces square-like artefacts, while the wavelet compression in JP2 produces a kind of blurred background. This blurred and smoothed background may be in case of manuscripts more disturbing than classical JPEG artefacts in the sense that it is less potent to create illusion of the texture of the original carrier (mostly paper or parchment) than the DCT. Even if we apply similar tests on higher quality digital photographs, we observe the disturbing blur in JP2 especially when wishing to save more space. Typically, if we wish to reduce the size of the output JP2 file twice face to the source JPEG file (e.g. 0.5 MB from ca. 1 MB on the 1200x1600 photograph), we will observe the dissolution and blurring of details with which we may be dissatisfied. In fact, if we wish to have similarly accepted results, there is no critical file size difference between JPEG and JP2. On the other hand, it is true that in case of wishing to compress the image very tightly, while having less quality requirements face to loss of information, the JP2 is a better solution than JPEG, but in the area of rather satisfactory results it seems that there is no real need to transfer images into JP2.

Of course, JP2 enables also a compression that might be rather interesting. Here, the only competitor might be only the PNG format in case we have tools that enable the set-up of the efficiency of its compression scheme (Graphic Workshop Professional, for example). PNG does not equal JP2, but the differences are not so much pronounced in comparison with TIFF/LZW or other solutions. As to our digitisation programme, it may be that we will come back in future to JP2 again, but for the moment the optimisation of our ca. 0.5 million images from manuscripts and similar materials for fast access remains in the JPEG domain, where the resolution and compression ratios are tuned for different types of originals. In conclusion, it may be said that for archival purposes, the classical JPEG format is being used, while for access optimised JPEG for manuscripts and old printed books, DjVu for digitised periodicals, and MrSID for digitised historical maps.

ACCESS SOLUTIONS

The digital data have been stored on CD-ROM for manuscripts and old printed books and in a mass storage library for digitised periodicals in which the data are written on magnetic tapes. The data quality control is based on regular measurement of physical and digital properties of optical media and on the self-control routine of the mass storage

system. In both cases, more copies are written on different media and stored separately. Even for the documents in the mass storage library - in which there are always two copies of each document - a third copy on magnetic tapes is stored off-line. The media problem can be solved only by a good monitoring policy, while the problems caused by changes of software and hardware platforms are anticipated by utilization of SGML (XML) based metadata containers and ISO data formats.

Another problem is, of course, the development of content descriptive formats that are mapped into SGML (XML) platforms. Here, modifications and various international agreements may change the scene more dramatically than media and format platforms themselves might require, because the desired data sharing may lead towards data migration, its completing and corrections rather faster than the changes of the encoding language, data formats, or applied media. This fact has been also one of the causes why we have decided to rewrite our Document Type Definitions after six or seven years since the establishment of our first structuring rules.

We also had to solve the practical problem of the relationship between archival facilities and on-line access applications. Initially, our conception was based on one storage facility used for both goals and assisted by some image server functions for access improvement. We introduced, for example, a DjVu on-the-fly conversion on demand directly from archival image files. However, the delivery from magnetic-tape based archival storage was slow, even if the speed of the on-the-fly conversion was not critical, as we dedicated a special server for this routine. We are aware of the fact that the fast development of large hard disk storage capacities could be a solution for this problem, but we had to fear other difficulties that were caused rather by vulnerability of on-line accessible digital archives from outside. After two hacker attacks, when we had to reinstall almost all our digital library systems, we decided to separate access data from the archival ones physically.

Thus, the development of two access systems has started this year: one for manuscripts and old printed books based on the shared catalogue of historical documents (Manuscriptorium) and one for other types of digitised documents, especially digitised periodicals (under development). Simultaneously, the migration of all the digital documents is taking place into new structures as well as development of new authoring and metadata structuring tools. In order to enlarge the number of co-operating institutions in the shared catalogue of historical documents, a tool is being developed to enable conversion of bibliographic records between MASTER format and UNIMARC (expected in the end of 2003). Our long-term goal is to create virtual research environments for study of our documents enhanced with selected full texts (TEI platform for manuscripts) and other useful information in electronic form. This is the theme of our new research plan for 2004-2010, which will be hopefully approved by December 2003. It is also the main reason for our participation in other national and international projects.

NOTES

1. For the W3C Schema see http://digit.nkp.cz/MMSB/1.0/msnkaip.xsd and its HTML documentation http://digit.nkp.cz/MMSB/1.0/dokumentace/msnkaip.html; while the root DTD is at http://digit.nkp.cz/MMSB/1.0/msnkaip.dtd - for parsers only – it consists of several autonomous entities/DTDs
2. See the complete definitions and English documentation at http://digit.nkp.cz/DigitisedPeriodicals/index_web.htm
3. See http://digit.nkp.cz/techstandards.html
4. Available for Linux and Solaris from http://djvu.sourceforge.net/
5. Eloquent samples from a beautiful Persian manuscript are at http://www.nkp.cz/start/knihcin/digit/vav/wavelet/Samples.html - in many parts – as to the quality parameters of various wavelet schemes - still valid is the study *Efficiency of Wavelet Compression* by Adolf Knoll available from http://www.nkp.cz/start/knihcin/digit/vav/wavelet/Efficiency-of-wavelet-conversion.html

REFERENCES

1. Knoll, A. *Digitization of Rare Library Materials: Storage and Access to Data* / Adolf Knoll … et al. Prague : National Library : Albertina icome Praha, 1997. 1 CD-ROM (Memoriae Mundi Series Bohemica). On-line in a shortened version, see: http://www.unesco.org/webworld/highlights/digitisation_0199.html or: http://www.nkp.cz/start/knihcin/digit/WWW/ENTER.HTM
2. Knoll, A. *New Image Formats and Approaches for Document Delivery and Their Comparison with Traditional Methods*. Published also on CD-ROM: Informace na dlani 2001. Praha, Albertina icome Praha, 2001. Paper presented at INFORUM 2001 Conference, 29 - 31 May 2001, Prague. http://www.inforum.cz/inforum2001/English/prispevky/knoll.htm
3. Knoll, A. *Compression of Bi-level Images: compressor performance report*. Published also on CD-ROM: Informace na dlani 2000. Praha, Albertina icome Praha, 2000. Paper presented at INFORUM 2000 Conference, 23 - 25 May 2000, Prague. http://www.inforum.cz/inforum2000/prednasky/kompresebitona.htm
4. Magnien, A & C. Arminjon. *Système descriptif des objets* mobiliers. Paris, Editions du patrimoine, 1999. 372 s. Documents et méthodes, 6 ; ISSN 1150-1383 ; ISBN 2-11-091636-2 ; ISBN 2-11-091765-2

WEB SITES REFERRED TO IN THE TEXT

Digitization of Rare Library Materials: Storage and Access to Data. http://www.nkp.cz/start/knihcin/digit/WWW/ENTER.HTM

DIEPER - Digitised European Periodicals. http://gdz.sub.uni-goettingen.de/dieper/

Kramerius. http://www.nkp.cz/o_knihovnach/konsorcia/VISK/VISK7.htm

Library Public Information Services Programme. http://www.nkp.cz/o_knihovnach/English/LPISindex.htm

LizardTech. http://www.lizardtech.com/

LuraTech GmbH. http://www.luratech.com/

Manuscriptorium. http://www.manuscriptorium.com/

MASTER - Manuscript Access through Standards for Electronic Records. http://www.cta.dmu.ac.uk/projects/master/

MASTER DTD. http://www.tei-c.org.uk/Master/Reference/DTD/masterx.dtd

Memoriae Mundi Series Bohemica. http://digit.nkp.cz/

National Library of the Czech Republic. http://www.nkp.cz/altnkeng.htm

RLG REACH Element Set for Shared Description of Museum Objects. http://www.rlg.org/reach.elements.html

UNESCO. http://www.unesco.org/

UNESCO's Memory of the World Programme. http://www.unesco.org/webworld/mdm/index_2.html