MANUSCRIPT DIGITIZATION AND ELECTRONIC PROCESSING OF MANUSCRIPTS IN THE CZECH NATIONAL LIBRARY

Zdeněk Uhlíř

Abstract: The paper informs about the history of manuscript digitization in the National Library of the Czech Republic as well as about other issues concerning processing of manuscripts. The main consequence of the massive digitization and record and/or full text processing is a paradigm shift leading to the digital history.

Keywords: manuscript digitization, processing of manuscripts, digital history, paradigm shift.

Introduction

More or less systematic digitization of manuscripts and other historical materials started only ten or fifteenth years ago. In the case of the National Library of the Czech Republic [3] it was in 1992 when the cooperation started within the UNESCO programme Memory of the World. The National Library of the Czech Republic has accomplished a great progress since these sheepish beginnings so that now its digitization team placed itself among the most advanced teams worldwide. The acquired experience during the twelve years shows that the large and massive digitization of manuscripts and other historical materials means a big challenge not only for the computer science and library and information science but for the history in large sense as well because it leads to the paradigm shift in general. This paper concerns some of these issues.

From Manuscript Digitization to the Digital History

As I have already said, the start of digitization in the National Library of the Czech Republic is dated in 1992 when cooperation was linked-up with the UNESCO programme Memory of the World. In 1993 the Czech National Library published a pilot CD-ROM [7] for this programme that was created – in close collaboration with the Albertina icome Ltd. [1] – as an example for similar activities in this framework. The prae-history of the digitization activities began at this point that consisted in learning from more developed and advanced teams and institutions as well as in gathering own experience. The results of this development released in 1995 were two pilot CD-ROMs of the newly-created Czech National Library's programme called Memoriae mundi series Bohemica, i.e. Antiphonarium Sedlecense [2] and Chronicon Concilii Constantiensis [4]. It was really the crucial experience leading to recognition that the biggest error concerning digital processing is to put together the data and the software. Thus, the main enlightenment following the first independent result was that the data and the software must be strictly and consequently divided. It is in a clear divergence from the old relational database tradition.

The effect of this recognition was creating the SGML based DOBM standard (Digitization of Old Books, Manuscripts, and Other Materials) that is a document type definition enabling makeing complex digital documents, i.e. compound documents of bibliographic and technical description as well as images-copies of the original documents. It enabled starting of a large and massive digitization for the programme Memoriae mundi series Bohemica. It is very important that following such an achievement on the one hand the programme Memoriae mundi series Bohemica became a national programme in 1998 and on the other the DOBM standard was accepted in 1999 as a UNESCO recommendation for its programme Memory of the World. The prae-history finished and the digitization in the National Library of the Czech Republic steped in its history.

History of digitization in the National Library of the Czech Republic consists in efforts at its webbisation because the net environment is the biggest challenge of the manuscripts work during last several decades. The first step in the effort was an attempt to make avilable via Internet a special text catalogue [9] that was inspired by the In principio in 1998, i.e. text catalogue created for a long time by the French Institute de Recherche et d´Histoire des Textes and the American Hill Monastic Manuscript Library. There were selected ten fields (library, shelf mark, leaves, rubric, first incipit, second incipit, explicit, structural ovwerview, bibliography and note-commentary) that should be able to characterise unambiguously each text item within the manuscript. Perhaps it was a good solution from the point of view of the content but problems came round the technical solution: the traditional relational database that was chosen however worked very slowly so that it was unacceptable for the final user. Through this negative experience it was discovered through this negative proof that another solution must be chosen. Such a solution consists in the use of markup language which enables representing not only a formal but also a content structure. In other words it means that all branches of manuscript work should be treated together or jointly. This was very important ascertainment.

Thus, in 1999 the National Library of the Czech Republic became one of partners of the European MASTER project (Manuscript Access through Standard for Electronic Records) [13]. The other partners were Humanities Computing Unit of the Oxford University, Center for Technology and the Arts of the De Montfort University Leicester, Institute de Recherche et d´Histoire des Textes in Paris, Dutch Royal Library at The Hague and Arnamagnaean Institute of the Copenhagen University. Representatives of various, perhaps quite different schools of manuscript work collaborated together in order to create an electronic record for manuscript descriptions. [12] The first idea was to prepare the records in SGML but the writing in XML appeared better at last. The new document type definition was ready by the end of the MASTER project in 2001 and it was widely disseminated through Europe. The importace of the MASTER DTD consists in two things: firstly, it enables preparing short and in-depth record using one document type definition only, and secondly, an extension of the descriptive record facing the complex digital document is possible. The National Library of the Czech Republic started useing it instead of the older DOBM standard because it is created exactly according to its long-term needs. The MASTER+ extension enabling connecting the manuscript record with the interrelated digital documents was done in 2002. History was brought to a close and the present started off.

Very important practical and organisational consequences followed after the creation of MASTER+ extension. As it makes possible creating whole complex digital documents, it enables building not only simple web presentations but a true digital library too. And digital library is not any ordinary resource, it is like a gate into the emerging virtual environment. Such a virtual environment must be understood in two simultaneous ways. Firstly, it is a net of newly originating institutions that are different and distinct from the traditional „stone“ institutions as we know them from the modern era. Such virtual institutions are in all probability consortia of traditional institutions that have some new goals, i.e. not only to preserve, conserve and lend the collection items but more likely to present the collection items in over-collection way and to re-present them from variuous points of view and in different sights. Thus, whichever virtual institution has different tasks in comparison to traditional „stone“ institutions. Secondly, it is a fluid compound representation of transient documents so that resource, not document appears at the first horizon. That means, not individual, but aggregate, collective phenomena are fundamental in such an environment. Very hard consequences follow within the information, communication, and knowledge sphere. Step by step a modern idea of objectivity is replaced by a new concept of virtuality. We are in the model of this great process and we do not know now what it will bring in the end but it is the biggest challenge of our present and near future.

Following these substantial ideas, the National Library of the Czech Republic initiated a decision to put together several of the most active institutions that take part in the Czech national project Memoriae mundi series Bohemica. Seven partners (apart from the Czech National Library also National Museum in Prague, Moravian Land Library in Brno, Research Library in Olomouc, Castle Library in Kynžvart, Museum of East Bohemia in Hradec Králové, Praemonstratensian Canony at Strahov) assembled and founded the Memoria project [14] in the end of 2003 that consists in collaboration in developing the virtual research environment for the work with historical holdings. Thus, the Memoria project is actually a consortium of institutions endeavouring to take step from the traditional information, communication, and knowledge environment into the virtual one, i.e. it is the genuine virtual institution. The Memoria consortium undertakes the Manuscriptorium database, [11] founded few months before in 2003 by the National Library of the Czech Republic and maintained by the Albertina icome Ltd. Manuscriptorium database is from its very beginnings oriented to the cooperation and integration with a wider circle of foreign partners. Universityy Library Bratislava in Slovakia became the first one. Other partners that signed agreement with the National Library of the Czech Republic as coordinators of the Manuscriptorium database on the manuscript cataloguing are the University Library Wrocław in Poland and the National and University Library Zagreb in Croatia. Currently testing of a more sophisticated integration is running. It consits in joining of the results of the Austrian project Monasterium [15] and the German one Codices electronici ecclesiae Coloniensis. [10] The Monasterium database this way represents an important attempt to integrate the cultural heritage at the supranational, transnational level for Central Europe.

Thus the main idea of the Manuscriptorium database is cooperation and integration. There are several aspects of this idea. Firstly, it concerns various organisational levels: the Czech National Library programme (Memoriae mundi series Bohemica); the Czech national programme (Libraries´ Public Information Services, branch 6); open group of the „willing“ partners within the Czech national programme (Memoria); and finally open group of the foreign partners (now Slovakia, Poland, Croatia, Austria, Germany – and the others wellcome: at the moment we are negotiating with the Library of the Lithuanian Academy of Sciences in Vilnius). Secondly, it concerns various material types (manuscript books, incunabula, early printed books, maps, graphics, charters, etc.), i.e. it is the idea of interdisciplinarity and transdisciplinarity that is the most important challenge of contemporary work with historical materials. Thirdly, it concerns integration of various document types (catalogue records, digital replicas/copies – images, sounds alike), fulltexts of primary, i.e. original, and secondary, i.e. interpretative, documents, eventually multimodal documents that are substantially compound and transient ones. And fourthly, it concerns integration of cultural heritage at the transnational Central European, eventually even European level. Thus, the main and proper goal of the Manuscriptorium database is to create a gate for the manuscript and other historical sources studies in a global dimension.

Now, it is very important to know how to do it in particular. The first step is to use the MASTER and MASTER+ records. Records created according to this double standard enable a goal-directed choice between the short records and on the other hand the in-depth records and subsequently to create a reasonable time-management. The MASTER records also enable the choice among various kinds of manuscript description according to the various purposes that the records are procured for and subsequently to aspire to interdisciplinarity and/or transdisciplinarity. Another important characteristic of the MASTER+ records is that they enable connecting the descriptive record with the appropriate interrelated digital documents and in this way making complex documents that are compound as well as transient, i.e. to build the digital library as a basis for the global virtual research environment. Last, but not least the use of MASTER and MASTER+ records is the necessary condition for interoperability because its´ consequence is the clear and apparent divide between the data and the software. Although to learn to create and to use the MASTER and/or MASTER+ records is a hard work the results of it are well-arranged and user-friendly.

The second step is to use standards for creating and processing images. These standards guarantee that the images will be of an excellent duality and subsequently that they can be archived and used again. The use of such standards enables making them accessible for browsing and/or searching according to the user´s purpose/s and subsequently building and providing an indirect service that is the keystone of the electronic, i.e. net and virtual environment. On the other hand, it enables building and providing a direct one which consists in digital reproduction delivery services. Such a posibility of combination of direct and indirect services is the greatest advantage of the electronic environment in comparison to the traditional environment of the printed book. In a further perspective using standards for creating and processing images leads to the posibility of building and providing various levels of indirect services according to the various quality levels that are made on the basis of various types of conversion, comprimation and so on and so forth. Of course, in this case there is some kind of authentication and licence management needed. It is a big challenge because contemporary copyright law as well as ideas about the intellectual property do not have at any case friendly inclination neither to the electronic resorces nor to the electronic environment.

The third step is to use a purpose targeted adaptation of the TEI standard. [17] It can be generated automatically using the Pizza Chef. [16] The TEI standard and its derivations enable creating various kinds of full texts concerning form of markup as well as content, i.e. it is able to process editions of the original documents on one hand and secondary documents (documents about these editions and/or documents about other documents, facts, events, persons, artifacts, etc.) on the other. Thus, the TEI standard follows the compatibility of all such documents and it enables subsequently very easy and comfortable archiving. The TEI document type definition makes possible a choice among various “markup ideologies” because of its flexible content based markup, i.e. among various even different purposes of the created document. That means not only the fulfillment of the interoperability requirements but also the possibility to do simple transformations and subsequently to make documents easily accessible and the possibility of a simple way to e-publishing. The consequence of compatibility and interoperability at this level is far-reaching. It means not only an implicit interdisciplinary and/or transdisciplinarity but also the possibility to evolve its´ explicitly. Upon the basis of only one archive database there are many presentation databases that can be created, maintained, and provided. The difference between individual presentation databases consists in the possibility to eke out the markup of document existing within the archive database with other specific markup according to the same or another TEI standard adaptation. It is a big advantage in comparison to traditional printed representations as well as a challenge to think in another ways than usual.

The fourth step is to use connectivity standards as Z39.50, OAI PMH alike. It enables integrating mutually various resources of the same document type, i.e. the primary documents (original historical documents and their various representations), the secondary documents (documents about primary documents and other interpretative ones), and the tertiary documents (catalogues and/or bibliographic records, documentation items alike) that use some kind of metadata (which stands to a reason nowadays). The consequences are the resource's interrelations and the completion of information. It enables creating indexes at the over-resource level, too. Such indexes facilitate heuristics within completed as well as individual resources. It is a welcome and important contribution for solving problems that so called second information crisis brought along. The connectivity protocols build an initial level of the virtual research environment because of easy orientation and navigation. Of course, even though it is very inchoate it is a step in the right direction.

The fifth step consists in the simultaneous, flexible use of various approaches. The most important thing is not to seek solitary ways but to use standardized and regularized tools of the net environment as Internet browsers, markup language editors and processors, parsers, validators, coding and format convertors alike. On the other hand together with the use of these regularized and standardized tools is quite a clear need to use and/or to create individual specific *ad hoc* applications as computational linguistics tools and systems etc. The multitasking facility of personal computer is another possibility how to do the work in the electronic environment more reasonably and user friendly because it enables using computer as a desktop in a quite real, not only metaphorical sense. Now, there are many and many such tools and applications at our command so that we can see entirely clearly that we need "new" methodologies instead of the "old" ones. Thus, although at the first sight it looks as if the easiest and simplest step according to all experience is the most difficult one. To seek "new" methodologies means herewith to develop a new paradigm, that is quite different from the old one, so it is a sure way to the uncertainty because it disrupts the previous certitude. Nevertheless if we go through this way step by step, we have learned and got to know that the paradigm shift comes slowly but surely.

The recognition follows that the issue of the virtual research environment for the work with historical holdings and cultural heritage is a question of its content, not of technics and/or technology. Thus, although technics and technology is *conditio sine qua non*, it is an insufficient condition. That means, as so as the information-communication technologies are an aid only and not the goal, the main problem concerns the content. The most important problem of the content is that the content presented as a cultural and scientific heritage is still the content of yesterday, not of today and tomorrow. So it must be adapted according to the requirements of the information wave and information society, not according to the ideas of the industrial ones. There is a crucial need to investigate which is the difference – we might say the *specific difference* – between these two fundamental conceptualizations, before it will be possible to develop any feasible and reasonable new ideas.

Of course, this problem is very large and difficult to be described or characterized in general. [18] So it must be narrowed for the domain of historical materials, i.e. cultural and scientific heritage only. The traditional conceptualization of cultural and scientific heritage within the industrial, i.e. modern society is based upon the idea of objectivity [8] of the world and of things within the world and subsequently upon the idea of the artifact. The idea of objectivity has been gradually replaced by the idea of intersubjectivity since only several decades ago. Artifact is not a simple and indifferent cultural object; it is a consciously and willfully created work. So the fundamental idea of cultural and scientific heritage is the idea of the work. The conceptualization of the work follows two main points of view, the historical and the philological. From the point of view of history *sensu largo* it is based upon the idea of external features in the sense of the *historische Hilfswissenschaften*. From the point of view of philology – and eventually also art history – it is based upon the idea of individuality and artificiality. The work according to the ideas of the industrial society is a real material thing, an external material object, not an ideal conceptual subject, an internal mental object. The work in this conceptualization simply is what is left, not it that it must be understood. The work in this sense is conceptualized as the *capta*, i.e. recorded and preserved data; it is not information in any meaningful sense. The understanding of cultural and scientific heritage within the industrial society is a consequence of that, a life in the industrial society is based on the operations with things, *atoms*. On the other hand, the arising information society is based on the operations with the signs and/or symbols, *bytes*. The difference between these two conceptualizations is a paradigmatic difference. Thus, because of different paradigms, i.e. different discourses there are different methodologies that construct different "facts". For the industrial and information wave/society the "facts" are not the same. That means, the "fact of yesterday" is not the "fact of today" and far less the "fact of tomorrow".

This is a real and difficult problem that must be solved if we want to go further. There are two natural ways how to solve this problem and they are given to us simultaneously. The first of them follows the recognition that objects, i.e. artifacts. Works are no external real objects but internal ideal objects; they are objects of human mind. They are some representations of external objects only. As they are representations they are herewith interpretations, explications, explanations, imaginations, fantasies, etc. Thus, as for cultural and scientific heritage, the paradigm shift doesn't concern as much preservation of the real objects as but rather preservation of such mental objects. That means that there is no need to preserve the all what was achieved but all what is apt for preservation. On the other hand, there is a question what it means, when we say "to be apt for preservation". If we agreed that cultural and scientific heritage is not solely the realm of things, i.e. not of the external objects, but more likely ideas, i.e. the internal subjects, then the preservation concerns ideas, not merely things. We must preserve especially ideas, not things, not the discourse concerning things and operating with ideas. As so as this discourse is the "discourse of yesterday" that operates with the "facts for yesterday". Thus, the goal of the work with historical materials, with cultural and scientific heritage is to create "facts of today" with the "regard for tomorrow".

The consequence of this recognition is that the path to the virtual research environment is not just a simple conversion from the traditional printed environment to the electronic one; it is not a transformation of printed representations to the electronic ones (and far less the mere retro-conversion). The firs main issue of the path to the electronic virtual environment is to find new forms of representations so that they diverge from the printed ones. The complex digital document (that is the essence of the virtual environment that consists in the evidence record connected with the interrelated documents) is no simple accumulation of individual partial documents but it is a structured net of documents that are transitional, i.e. they generate the integrated transitional compound document. [21] This could mean that such a complex document doesn't always need to have the same constituents and that the constituents can differ in various request situations. If so, virtuality is given at least potentially. The question is how to bring potential virtuality into the actual one. Thus, the other issue of the path to the electronic virtual environment concerns methodologies and facts created according to the new methodologies, i.e. it is the issue of paradigm shift. The question of the methodology as well as paradigm is the crucial question now.

The first substantial step during the way to the virtual research environment for the work with historical holdings and/or cultural and scientific heritage must be a change of understanding of cataloguing, bibliography, documentation, etc. The previous understanding of these activities and subsequent products is based on the typical situation of printed publications because it comes out from the external features and the idea of the text as work. It does not correspond with the typical situation of manuscripts and dominant manuscript environment. It does not correspond with the typical situation of electronic or digital documents. [20] Now such an understanding of cataloguing, bibliography, documentation, etc. is crucial which is oriented to the internal features in the sense of the *historische Hilfswissenschaften* [19] and to the idea of the text as floating continuum, [3] not as the work. Requirements for bibliographic and other similar records deviated from the manifestations (publications, editions) and items (copies, holdings of publications) to expressions and works in a quite virtual sense. [5] Subsequently there is a real possibility that the same (i.e. identical) item as well as manifestation (text, document) can be part or constituent of various expressions and works so that we must accept a new type in these scales, the work expression that is not simple accumulation of the expression and the work. It is a fully new domain of knowledge that must be seriously searched.

Conclusion

Thus, the digitization team of the National Library of the Czech Republic asks what to do in the near future. There are three fundamental tasks: first of all, transferring all the catalogues, bibliographies, documentations, factual and material studies as well as historical text editions into the virtual, i.e. electronic, digital environment; second of all, creating many information communication technologies tools for the mass processing of historical documents and holdings; and third of all, starting creating of the multimodal resources for presentation and re-presentation of the integrated cultural and scientific heritage. It will take some years – and then we will see furthermore.

Bibliography

[1] Albertina icome, see at the URL http://www.aipberouon.cz.

[2] Antifonář Sedlecký. Antiphonary of Sedles. Antiphonarium Sedlecense, MS XIII A 6, ed. Zdeňka Hledíková – Hana Hlaváčková – David Eben. CD-ROM, Praha, Národní knihovna & Albertina icome Praha, 1995.

[3] Bryant, John: The Fluid Text, Ann Arbor, 2002.

[4] Chronicon Concilii Constantiensis. Malá Riechentalova kronika, MS VII A 18, ed. Zdeněk Uhlíř. CD-ROM, Praha, Národní knihovna & Albertina icome Praha, 1995.

[5] Functional Requirements for Bibliographic Records: Final Report, München, K.G.Saur, 1998, available at the URL http://www.ifla.org/VII/s13/frbr/frbr.pdf .

[6] National Library of the Czech Republic, see at the URL http://www.nkp.cz.

[7] Paměť světa. Mémoire du Monde. Memory of the World. CD-ROM, Praha, Národní knihovna v Praze & Albertina icome Ltd., 1992.

[8] Popper, Karl Raimund: Ausgangspunkte. Meine intelektuelle Entwicklung, translated Friedrich Griese, München, Piper Verlag, 2004; Popper, Karl Raimund: The Open Society and Its Enemies, London, Routledge & Kegan Paul, 1962; Popper, Karl Raimund: The Poverty of Historicism, London, Routledge & Kegan Paul – Boston (Mass.), The Beacon Press, 1957; Popper, Karl Raimund: The Logic of Scientific Discovery, London, Hutchinson & Co. – New York, Basic Books Inc., 1959.

[9] Text Catalogue, see introduction available at the URL http://digit.nkp.cz .

[10] The Codices electronici ecclesiae Coloniensis project, available at the URL http://www.ceec.uni-koelnn.de/, eventually http://www.ceec2.uni-koeln.de/.

[11] The Manuscriptorium database, available at the URL http://www.manuscriptorium.com.

[12] The MASTER document type definition, available at the URL http://www.tei-c.org.uk/Master/Reference/DTD/, eventually http://www.tei-c.org.uk/Activities/MS/FASC-ms.pdf.

[13] The MASTER project, available at the URL http://www.cta.dmu.ac.uk/projects/master.

[14] The Memoria project, available at the URL http://www.memoria.cz.

[15] The Monasterium project, available at the URL http://www.monasterium.net.

[16] The Pizza Chef tool, available at the URL http://www.tei-c.org/pizza.html .

[17] The TEI document type definition, available at the URL http://www.tei-c.org/Guidelines2/index.html.

[18] Toffler, Alvin: The Thire Wave, London, Pan Books, 1981; Negroponte, Nicholas: Digitální svět. Being Digital, translated Petr Koubský, Praha, Management Press – Softwarové noviny, 2001; Flusser, Vilém: Do universa technických obrazů, translated Jiří Fiala, Praha, OSVU, 2001; Flusser, Vilém: Kommunikologie, ed. Stefan Bollmann – Edith Flusser, Frankfurt am Main, Fischer Taschenbuch Verlag, 2000.

[19] Uhlíř, Zdeněk: Teorie a metodologie elektronicko-digitálního zpracování rukopisů a hybridní knihovna, Praha, Národní knihovna České republiky, 2002.

[20] Uhlíř, Zdeněk: Terminologie a pojmy v čase paradigmatických změn, Národní knihovna: knihovnická revue, 14, 2003 (4), p. 236-244.

[21] Williams, Robert F.: What´s New: Transient Compound Documents Establish Irrevocable Records, available at the URL http://www.cohasset.com/main/library/coh_articles/whatnew_body_transient.htm.

Author Information

Zdeněk Uhlíř – National Library of the Czech Republic, Department of Manuscripts and Early Printed Books, Klementinum 190, 11000 Prague, Czechia; e-mail: Zdenek.Uhlir@nkp.cz