УДК 811.362.3

A. Knoll

National Library of the Czech Republic

MANUSCRIPTORIUM: SEAMLESS ACCESS TO OLD EUROPEAN WRITTEN HERITAGE

The article has to do with the EU project «Manuscriptorium», its principles, its usage, its problems and its future. Deep attention was given to correct production of digital data and search of appropriate approaches to ensure a relative independence of concrete computer environment. It was mentioned an agreement with UNESCO to support the new Memory of the World programme with digitization of relevant materials to demonstrate the strength of this technology. They were accentuated especially in the moment when we decided to pass from small discrete projects to routine digitization of manuscripts. It was in 1995 and it made us to think about a format that should contain the entire digital document: both metadata and data. It was realized the external character of binary data face to descriptive texts and from here only a short way led to the building of a compound digital document via mark-up languages.

To aggregate data from partner digital libraries into a seamlessly unifying application the first format passage took place due to National library willingness to create and launch a digital library. Building of Manuscriptorium as an international aggregator for digitized historical documents (mostly manuscripts) has been reviewed. The solution is said to be based on a central database that collects metadata records from partner libraries.

Key words: manuscriptorium, seamless access, archive, digitization, discreet.

Кнолл А., Національна бібліотека Чеської Республіки. ***МАНУСКРИПТОРІУМ: ЄДИНИЙ ДОСТУП ДО АРХІВНИХ ЄВРОПЕЙСЬКИХ ДОКУМЕНТІВ***

Розкрито аспекти проекту «Манускрипторіум», що уможливлює легкий доступ до інформаційних ресурсів, у тому числі письмових архівних документів, цінних для наукових досліджень у сучасній гуманітарній парадигмі. Виділено та проаналізовано деякі проблеми щодо технічних можливостей користувачів, а також аспекти використання єдиного електронного доступу до архівних документів.

Ключові слова: манускрипторіум, єдиний доступ, архів, дигітація, дискретний

Кнолл А., Национальная библиотека Чешской Республики. ***МАНУСКРИПТОРИУМ: ЕДИНЫЙ ДОСТУП К АРХИВНЫМ ЕВРОПЕЙСКИМ ДОКУМЕНТАМ***

Раскрыты некоторые аспекты проекта «Манускрипториум», облегчающего доступ к информационным ресурсам, в том числе редким архивным документам, представляющим интерес для ученых в парадигме современных исследований. Внимание уделено техническим аспектам проекта, а также возможностям единого доступа к ресурсам для пользователей.

Ключевые слова: манускрипториум, единий доступ, архив, дигитализация, дискретний.

In 2002, the National Library of the Czech Republic launched the Manuscriptorium Digital Library that completed a period of our efforts to enable digital access to old manuscripts and, simultaneously, started a new one.

In the pre-Manuscriptorium period, a lot of attention was given to correct production of digital data and search of appropriate approaches to ensure a relative independence of concrete computer environment. The very starting point was our agreement with UNESCO to support the new Memory of the World programme with digitization of relevant materials to demonstrate the strength of this technology. In those early years,

not only a lot of optimism and strength were shown, but almost immediately also its weaknesses began to appear. They were accentuated especially in the moment when we decided to pass from small discrete projects to routine digitization of manuscripts. It was in 1995 and it made us to think about a format that should contain the entire digital document: both metadata and data. We realized the external character of binary data face to descriptive texts and from here only a short way led to the building of a compound digital document via mark-up languages.

We started with our own SGML solution in 1996, passed to a more complex XML environment in 2002 and finally arrived to a TEI P5 application in 2009 [1–3]. Whilst the first format passage took place because of our willingness to create and launch a digital library, the second passage was already a joint international endeavour that foresaw our preparedness to aggregate data from partner digital libraries into a seamlessly unifying application. This upgrade was possible thanks to the EU ENRICH project that aimed at further building of Manuscriptorium as an international aggregator for digitized historical documents, mostly manuscripts.

The solution is based on a central database that collects metadata records from partner libraries. These metadata records must contain bibliographic descriptions and structural maps that enable referencing to existing data representations of structural elements (mostly pages) on partner servers. The compound document is thus created during the user's session from available metadata and remote images or text files (in case that the data representation is full text). The user makes various systems to cooperate in real time to get his document in its complexity or to create his own collections or virtual documents.

From the users' point of view, this service makes possible to gather dispersed rare collections in one place that is a great advantage for research, because the traditional problem for users was to travel physically in space to get access to various documents they needed for their research, and even in case of available digital surrogates, the physical travel this time became a virtual one from one application to another facing different behaviours, rights, tools, or having to cope with various different opportunities.

Manuscriptorium brings a good solution for this with its distributed model of a digital library. Usually the solutions are portals such as various information gateways (e.g. TEL or Europeana) or centralized digital libraries as it is the case of the World Digital Library. While the model of a centralized digital library is good and appropriate for a single producer or a group of producers sharing the same formats and at least partly the same infrastructure, for seamless aggregation of heterogeneous resources the distributed model is better, because it enables continuous acquisition of partners' data through communication protocols (OAI).

The internal TEI P5 Manuscriptorium formati can accommodate both library MARC-based descriptions and scientific TEI descriptions. This is an advantage over the older document format. Furthermore, for Czech institutions that digitize their documents within the framework of a national programme supported by the Ministry of Culture, the National Library of the Czech Republic guarantees also long-term storage of produced data. For this additional metadata and naming and numbering conventions are prescribedii. Our philosophy is to accept partner data as they are, i.e. that even if the OAI harvest is preferred we can process data offered off-line provided that they contain descriptions, structure, and the partner offers JPEG images (also PNG or GIF) to be called into the Manuscriptorium interface during the user's session.

We support not only transformations of existing data, but we also offer tools for creation of TEI P5 compatible documents. This is done via the on-line M-TOOL appli-

cation that helps creation of descriptive and structural metadata with links to existing images that represent the document pages or folios. The registered user who creates the digital documents via the above-mentioned application can also work in the so-called on-line Manuscriptorium for Candidates (M-CAN) application that serves for verification of created documents and enables dialogue with Manuscriptorium supervisors and its technical administrator up to upload of correct records and inclusion of such documents into Manuscriptorium.

The integration of M-TOOL into Manuscriptorium is, however, deeper than this one: it enables also creation of virtual documents on the basis of users' selection of individual pages across the whole Manuscriptorium independently of physical location of images. Such documents can be also shared with others. This is already about Manuscriptorium personalization opportunities that are contained in users' space called my library. This space can contain, besides the virtual documents, also users' collections complying with their refined search criteria (dynamic collections) or just consisting of individually picked-up items (static collections). The advantage of the dynamic collection is that the stored user's query returns always the results based on a new search, i.e. in case that new items complying with the search criteria have been added, the user's collection is automatically enriched.

We expect to launch a new version of Manuscriptorium towards the end of 2013 in which we hope that all these described features will be improved.

The metadata of Manuscriptorium documents are shared via communication protocols with other aggregation services of local (Czech Uniform Information Gateway), but especially of European or world importance and impact. These are the both cultural and professional services: thus, on one hand Manuscriptorium cooperates with The European Library TEL, EUROPEANA, and CERL-MSS portal, while on the other hand it is indexed by various resource discovery services operated by the companies such as EBSCO, SUMMON, ExLibris, or recently SUWECO. This enables even to small institutions to be immediately visible on a world scale.

As of 22 August 2013, Manuscriptorium contained 305,855 records from which 24,844 were those of fully digitized documents, while 605 from them had also full text representations. As to fully digitized documents, about 75 % are from other countries than from the Czech Republic. The biggest contributor is still the National Library of the Czech Republic that had in Manuscriptorium 3,320 items as of December 2012, while the other biggest Czech contributors were the Moravian Library in Brno with 470 items, Strahov Monastery Library (319), and the National Museum Library (272) – altogether Manuscriptorium has 51 Czech partners among research libraries, museums, archives, monastery libraries, and others, as of September 2013.

The importance of foreign cooperation and the international dimension of Manuscriptorium are well demonstrated by listing the most important foreign partners: Complutense University Library in Madrid (Spain – 2902 items), St. Trinity Sergiev Monastery in Sergiev Posad (Russia – 2668), Wroclaw University Library (Poland / 1839), Collections administered by the Cologne University Library (Germany – 1634), National Library of Italy in Florence (Italy – 1566), National Library of Spain in Madrid (Spain – 1444), National and University Library of Iceland together with Arne Magnusson Foundation in Reykjavík (Iceland – 1176), Vilnius University Library (Lithuania – 1085), Heidelberg University Library (Germany – 1025), e-Codices Digital Library (Switzerland – 889), National Library of Romania in Bucharest (Romania – 393), University Library of Bratislava (Slovakia – 241), and Zielona Góra University Library

(Poland – 231) – altogether we have data from 75 foreign libraries from which 27 are from the Swiss e-Codices project. A Psalm Book from the Ivan Vazov Public

There are also a lot of data that can be taken from the web statistics. Here we can see how important the role of EUROPEANA as Manuscriptorium traffic generator is in last three years. From all the visits, EUROPEANA is on the 3rd place with 13.93 %, preceded only by direct visits (23.47 %) and Google (21.78 %), while the other important traffic generators are the National Library of the Czech Republic itself (5.95%) or other domestic or foreign partners, but also the most important Czech search engine (Seznam – 3.58 %), Czech Wikipedia (6th place – 2.58 %), or Facebook (9th place – 0.80 %); for reference, The European Library TEL is on 16th place with only 0.49 %. However, when we take referencing pages, EUROPEANA is the leader with 27.50 %, followed by the National Library of the Czech Republic with 11.57 % and Czech Wikipedia with 5.11 %, while Facebook still being rather important on 6th place with 1.59%. We are happy especially seeing how important and fruitful the Manuscriptorium inclusion into EUROPEANA has been. Perhaps thanks to this, we can also observe the growth of the share of direct visits (+6 %) during the period when EUROPEANA started to play such an important role for us as a traffic generator. This means that an important number of users who went once to us through EUROPEANA began to enter Manuscriptorium directly, becoming thus our new users. As to users, slightly more than a half are from the Czech Republic, while the three first foreign countries are stable in order of number of visits since 2009: Germany, Poland, and U.S.A., followed by Italy, Spain, Austria, France, Romania, and Slovakia. Here we can see over time the growth of interest especially in Italy, Spain, Austria, United Kingdom, and Russia. Altogether Manuscriptorium was accessed in the period July 2012 – August 2013 by visitors from 161 countries. As to non-European countries, important numbers of visits came – besides U.S.A. – from Canada, Mexico, Australia, Japan, Brazil, Israel, Argentina, but also from India, Morocco, Saudi Arabia, Egypt, Philippines, or Iran. From here we can see how important is to join together dispersed collections in the virtual Manuscriptorium world. Even if – with the exception of the National Library of Kazakhstan – all Manuscriptorium partners are from Europe, their collections contain also the documents that are relevant for other cultures.

Geographical provenance of Manuscriptorium users (Sept 2012 – Aug 2013)

Another interesting point of view consists in comparison of usage with other electronic resources offered by the National Library of the Czech Republic especially with licensed electronic resources. Even if Manuscriptorium has fewer visits, the number of accessed pages is roughly identical. In comparison with another digital library of ours, that for modern materials – periodicals and monographs published after 1800 called Kramerius – Manuscriptorium has slightly more than one third of Kramerius' users, but, of course, Kramerius has seven times more of accessed pages, because its materials are understandable and generally readable. When we take into account how special Manuscriptorium is, these usage results are very positive. We have several known problems that may be ordered in two groups: 1. technical and organizational and 2. political and cultural problems.

The first group of problems depends on partners' cooperation and their reliability: in some cases the partner servers do not always function, permanent URLs of images have been changed without update of the OAI harvested profiles, and of course, better funding for aggregation of documents would be also necessary especially for supporting inclusion of new partners, i.e. funding available partly for them and for us.

The second group of problems contains our hesitation about how far it is feasible to include documents from Eastern Asia (e.g. Chinese or Korean), whereas documents from Near East (Arabic, Persian, Ottoman, etc.) are already included at least from European collections. It is also true that some people, countries, or institutions may dislike to be aggregated by a service operated by a Czech institution and also some people are still unwilling to make their collections widely accessible.

Another technological and cultural problem is inclusion of documents written in other characters than in Latin alphabet, because the metadata of such documents are usually in original characters when the partner is from the country where the language is or was spoken or used, while for similar documents from other countries, the titles and the metadata that should remain in the original language are usually transliterated… and we may have several different transliterations as we have for example for Cyrillic documents. In practice, we already have for them transliterations following Middle-European rules (using Czech diacritic signs), Anglo-Saxon/American rules, and even local national transliterations as e.g. the rules used in Romania for transliteration of the Cyrillic script. This fact makes problems when searching such documents and it may be only partly solved by creating special virtual collections. Of course, another problem is usage of different levels of individual languages that creates spelling problems, because, for example, the 14th century Czech is a different language from the Czech used in 18th or 20th/21st centuries that we may use for descriptions.

Today, Manuscriptorium is funded only by the National Library of the Czech Republic thanks to understanding of the Czech Ministry of Culture; we also enjoy of some funding for research and development, but in principle all this money enables only survival and slower development than we would wish. In spite of this, we may expect a new version to appear by the end of 2013. It will be more user-friendly and the way how it will look like is based on collection of users' feedback.

Internally, we have been working on pilot solutions for inclusion of external thesauri to improve the search possibilities, mark-up of music documents, pattern recognition in images, automated recognition of handwritten texts, comparison of full texts, etc.

Even if wishing to make faster progress especially in the area of further aggregation than we are able to, we think that Manuscriptorium has become a useful tool for research of historical documents.

References

1. **Knoll A.** Projects / A. Knoll. – Asses mode : http://projects.oucs.ox.ac.uk/ENRICH/
2. **Knoll A.** Manuscriptorium principles / A. Knoll. – Asses mode : http://www.manuscriptorium.com/sites/default/files/docs/manuscriptorium_visk6_definice.pd
3. **Knoll A.** Manuscriptorium Project / A. Knoll – Asses mode : http://www.manuscriptorium.eu

Надійшла до редколегії 05.02.14