Národní knihovna
České republiky
National Library
of the Czech Republic

Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	Mgr. Adolf Knoll	
Pracoviště – dle organizační struktury	Tajemník pro vědu, výzkum a mezinárodní vztahy	
Pracoviště – zařazení	KGŘ	
Důvod cesty	Účast na konferenci QQML 2012	
Místo – město	Limerick	
Místo – země	Irsko	
Datum (od-do)	21. – 26. května 2012	
Podrobný časový harmonogram	21. května 2012 – přílet ve večerních hodinách 22. - 25. května 2012 - konference 26. května 2012 – odlet brzy ráno	
Spolucestující z NK	xxx	
Finanční zajištění	Institucionální výzkum úkol 0135	
Cíle cesty	Účast na konferenci na základě přijetí referátu.	
Plnění cílů cesty (konkrétně)	Hlavním cílem cesty bylo především přednést referát na téma rozvoje Manuscriptoria především v oblasti bezešvé agregace dat, příp. získat další obsahové partnery z dalších kulturních institucí, a seznámit mezinárodní odbornou veřejnost s výsledky VaV v této oblasti.	
Program a další podrobnější informace	http://www.isast.org/programabstractregister.html text vystoupení viz příloha	
Přivezené materiály	xxx	
Datum předložení zprávy	30. května 2012	
Podpis předkladatele zprávy		
Podpis nadřízeného	Datum:	Podpis:
Vloženo na Intranet	Datum: 30. května 2012	Podpis:
Přijato v mezinárodním oddělení	Datum:	Podpis:

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

Homogenizing Access to Heterogeneous Resources of Digitized Manuscripts

Adolf Knoll

National Library of the Czech Republic

Abstract: The Manuscriptorium Digital Library is seamless aggregator of data from ca. 100 memory institutions in Europe. It offers tools for data contributors as well as for user-oriented personalization. The metadata from Manuscriptorium are used by many other services, such as The European Library, EUROPEANA, or SUMMON and EBSCO resource discovery services.

Keywords: digitization, data aggregation, seamless access, TEI, structural maps, personalized access

1. Introduction

In 2000, a new standard for electronic access to manuscripts was prepared within a European project in which also the Czech National Library had taken part. It was defined on the TEI platform and called MASTER, but it solved only the descriptive part of the complex digital document; therefore, it was enlarged for the needs of our digitization project that was transformed in 2000 into a national grant programme supported by the Ministry of Culture. The enlargement consisted in adding elements for description of the structure of digital documents and the technical information concerning the produced digital images – this part was taken from the document called Data Dictionary for Still Digital Images that had at that time the status of draft of a planned NISO standard[1] and the DIG35 specification[2]. The final standard was known as *masterx* standard, more precisely as msnkaip.xsd[3]. It was used until 2009 when it was replaced by another, this time fully international, standard, the so-called TEI P5 ENRICH schema[4].

[1] ANSI/NISO Z39.87 - Data Dictionary - Technical Metadata for Digital Still Images. Bethesda, NISO Press, 2006. 107 pp. ISBN 1-880124-66-1
[2] DIG35 Specification: Metadata for Digital Images, version 1.1. See: http://www.i3a.org/resources/#metadata
[3] http://digit.nkp.cz/MMSB/1.1/msnkaip.xsd
[4] http://www.manuscriptorium.com/index.php?q=content/tei-p5-enrich-sch%C3%A9ma

2. Data production

The first steps in digitization of old library materials were done in 1992/1993, while since 1996 it was a routine job. Thanks to two national grant programmes that appeared in 2000, dozens of Czech institutions could make many items from their collections accessible. The programmes were built on de facto national standards and the data were made accessible in 2003 in two digital libraries (Manuscriptorium for manuscripts and old printed books and Kramerius for digitized periodicals and modern monographs), but the work was always far away from what could be called mass digitization: from the collections of the National Library ca. 10 million pages were digitized in this way, a greater part being data reflecting documents published after 1800. To this, it is necessary to add other institutions that also provided access to selected important library documents.

The necessity to speed up the data production was felt both in the domain of old and modern documents. Fortunately nowadays, two big digitization projects have started to cover hundreds of thousands of original documents. Whilst for modern documents it is a project making use of EU structural funds, for historical collections, the relevant activity is the Google project whose aim is to digitize ca. 200,000 volumes of historical printed books (mostly 17^{th} and 18^{th} centuries) until the year 2017. The data from this project will be accessible from Google services as well as from the digital services of the National Library, as there is no limitation in use provided the data will be accessible freely for non-commercial services. The open free access is the fundamental Google condition for entering in cooperation with the library.

In parallel, digitization of manuscripts, incunabula, and very rare old printed materials is going slowly ahead as national and institutional funding makes it possible.

3. Adding value through digital library

Digitization gives users possibility to visit many collections from one place; this is crucially important especially in the case of unique historical documents and namely manuscripts, because the documents even if created in the same scriptorium or territory, were travelling in space and are now dispersed in various collections in the whole world.

This fact inspired us to work on virtual aggregation of such documents. We wished to overcome the limitations given by portals where you can find what you need, but to see it; you have to travel again, not physically this time,

but anyway virtually. In practice, it is about consulting various different systems with various access rules and user behaviours.

The structural maps in complex (compound) digital documents, however, make it possible to browse books over space provided we know where the referenced data files (mostly images) are stored and provided they are always available through http protocols on their concrete URLs.
This was the basic idea on which we started to build a new Manuscriptorium Digital Library as a place where all the different data from various systems and repositories can be presented to users in an homogeneous manner as if all of them were coming not only from the same database, but also from the same image bank. The truth, however, is that in real time the image banks of all the Manuscriptorium partners are being used in function of the work done by users via the Manuscriptorium interface.

The way to achieve at least partially this goal was not easy. There was a period of time in the first years after the launch of Manuscriptorium in 2003, when we were registering the interest of several other foreign institutions to share their metadata in the same database in order to create a larger platform for the work with the common cultural heritage. The first institution from outside of the Czech Republic was the Wroclaw University Library in Poland. Step by step it was also realized that not only sharing the metadata, but also the data could be very useful, because this would facilitate the research if this could be done under the same interface and same tools independently of the primary access tools and interfaces in the partner digital libraries.

However, to make it real, the willingness of institutions to take part in such cooperation had to be supported by a common project to speed up the work. This was done thanks to the EU ENRICH project (November 2007 – November 2009).

3.1. The main tasks of data aggregation

The main tasks were to aggregate content from partner institutions and to provide new tools for users. However, the situation was more complex, because the partners and other content institutions worked under different conditions: some of them had digital libraries and some of them had not in spite of having digital content eligible for aggregation; therefore, a more thorough mapping of their possibilities and needs had to be done.

3.1.1. The partners

The final target was to achieve seamless presentation of partners' data. For this, it was necessary to have at least the catalogue/bibliographic descriptions of the items to be incorporated into the Manuscriptorium Digital Library as well as their structural maps containing references to visual representations of manuscripts located in remote partner repositories/digital libraries. The optimal solution was to harvest these metadata through OAI-PMH protocols.

The formats of (bibliographic) descriptions expressed the approaches of the two communities involved in the work: those who administered the collections and those who made research on these collections. Typically, the first group used mostly the MARC-family formats, while the second group was using TEI platform for description and also transcription of the works. It is also true that some collections owners were using also TEI-oriented solutions.

The discussion was also about where the transformation of the original records should take part:

1. during ingest of records with application of the Manuscriptorium internal format as a common base for all further handling of metadata;
2. only in the moment of extraction of metadata for indexation and display. The latter approach supposed application of the METS container that would preserve the original format of the descriptive metadata.

At that time, the Manuscriptorium internal format was the TEI P4 MASTER (enriched with structural metadata) and it was used also by some partners, while simultaneously the application of TEI P5 was considered by several other partners. This split also the TEI community in two. In the contrary, the METS approach could be politically more acceptable, because it required from partners neither any adaptation of the descriptive format nor it supposed any winning solution to be applied for resulting metadata transformations. Surprisingly, the international consortium decided to write a new TEI P5 definition for description and structuring of digitized manuscripts. The result was a new schema as shown more above. Furthermore, there was a unanimous requirement to build this schema in Manuscriptorium as its internal format. This made application of METS unneeded, because the new TEI P5 schema contained all the necessary elements for description and structuring, while its granularity was on the level to accommodate not only very detailed

requirements of researchers, but it showed also the necessary MARC-family compatibility.

Thus, the new international format for description and structuring of manuscripts was created. As most TEI detailed applications, it presented problems to newcomers, because being much larger than usual library MARC records. In order to facilitate creation of new TEI files, an online application was developed that enabled preparation of necessary (even very analytically detailed) descriptions and structures. It is available at the Manuscriptorium site as M-TOOL ONLINE[5]. It was built on the experience gathered around usage of its standalone predecessor that supported the previous *masterx* schema. Its usage is combined with another tool for validation and control of produced TEI P5 (enrich.dtd) compliant files. This tool is called M-CAN [6]and it builds on the experience of the previous Manuscriptorium for Candidates virtual space where the partners can upload the files that are assigned similar behaviour as in the real Manuscriptorium. From here, the partners can offer the files for further processing aiming at aggregation of the described manuscripts into Manuscriptorium.

All the metadata are indexed in Manuscriptorium database and during users' work with the digitized manuscript the image files are accessed directly from remote partner servers. This is the same also in case of receipt of partner metadata via OAI or in other ways, e.g. offline in batch for further processing.

3.1.2. The users

Primarily, the users do not get only aggregated metadata that enable various search strategies based on an homogenized index, but in comparison to other aggregation initiatives from various resources placed in a heterogeneous virtual environment they are getting also a uniform access to the virtual representation of manuscripts, i.e. their behaviour is only one independently of the requirements of all the external systems that contribute with their data to Manuscriptorium.

In this way, the users can be offered a virtual space within Manuscriptorium that they can personalize as they wish. The personalization tools enable creation of virtual collections and virtual documents:

---

[5] See: M-Tool – User's Manual. Version 2.0 online at http://www.manuscriptorium.com/apps/m-tool/docs/m_tool_manual_en.pdf

[6] See: Manuscriptorium for Candidates Help File at http://www.manuscriptorium.com/apps/candidates/help/help_mcanp5_eng.pdf

- The virtual collections contain only the items that are of interest to the concrete user. They can be static or dynamic. Whilst the static collection is a mere list of selected items, the dynamic one is the result of application of a search query. In fact the query formula is stored in the user account after having been tuned by the user until it satisfies its needs and when he comes back, it is applied again on the database index. In case this one has been enriched with relevant items since the last user's visit, the user is returned his virtual collection enriched with all the newly added manuscripts that comply with the search query. Thus the dynamic collection is growing together with the entire digital library.
- The virtual documents are the documents that do not exist in reality, because they are created by the user from any analytically available digitized pages in Manuscriptorium and also from those freely available on Internet. In this way he can create a virtual book consisting of folios or pages that correspond with his needs even if they come from the documents dispersed geographically in space and also virtually coming from different digital libraries, repositories, or services. For creation of virtual documents, users work with the M-TOOL ONLINE that enables them to describe the newly created items in conformity with the TEI P5 schema; they are also assisted by the Manuscriptorium viewer that supports a very easy selection of the pages required for incorporation into the virtual document. This one gets the same behaviour as any other really existing Manuscriptorium document.

4. The plans for the future

Nowadays, Manuscriptorium (http://www.manuscriptorium.eu) has ca. 100 contributing partners, from which a half is from outside of the Czech Republic. Thus, Manuscriptorium is an important data aggregator for the EUROPEANA portal and, of course, for The European Library (TEL). It has already brought into EUROPEANA data of a considerable quantity of institutions that would not have otherwise the capacity to do it themselves.

4.1. Metadata and texts

As the descriptive metadata (and also full texts if applicable) come from different cultures and periods of time, there is a large linguistic variety in Manuscriptorium due not only to a big quantity of descriptive and content

languages, but also due to various stages of their historical development. To this, we should add different ways of transcription of non-Latin characters thanks to different rules applied in participating countries. In this way, we have, for example, Slavonic manuscripts described in Cyrillic and the original language as well as in various Latin transcriptions (Czech/Central European, English, Romanian, etc.) and local cataloguing languages. This creates also a large variety of terms designating the same physical persons, places, or events.

To bridge the problems during the search, the Manuscriptorium team decided to test application of two approaches: to deal with variants of graphemes and to use external thesauri for a better relevance of searches:

- Graphemes are the smallest semantically distinguishing units in a written language. Various different graphemes can be used to express the same phoneme (or the phoneme interpreted as being same from the point of view of another different linguistic environment). In such a way, for example, the graphemes <a> and <á> are phonologically relevant in Czech, but not relevant in some other languages, and we may imagine how the things become complex in a multilingual environment containing in plus different stages of evolution of individual languages and different and frequently erroneous spellings of the same word especially in historical documents when the level of application of unified orthography was low. For this reason, Manuscriptorium offers the option to avoid possible problems through application of techniques dealing with the problems related to graphemes.
- The first external thesauri applied in a pilot project concerned names of places: we used the Czech thesaurus called *Old Towns (Stará města)* and the CERL Thesaurus of names of places. The results still need to be evaluated more before taking the decision to implement them into operational use.

4.2. Imaging

When the metadata are in the central database, the images are spread over different remote repositories. Manuscriptorium works only with the image formats recommended for the World Wide Web, i.e. JPEG, GIF, and PNG, while special tile techniques are used for representation of large data files generated especially from digitization of historical maps. For processing of these data, we need to store them in our repository, but the number of maps in

Manuscriptorium is rather low so that the main concern is about the data representing volumes of manuscripts and rare printed books.

When thinking about possible further work with images, we concluded that it would be positive to be able to recognize automatically various content elements and various typical shapes or objects in images. On a pilot level, we have been testing recognition between some types of pages: those containing illustrations, painted initials, notations of music, and comments on margins, tables, or bordures. To be able to do this, the images have to be pre-processed locally to generate necessary metadata. Thus, this feature is applicable only to the data stored in the central Manuscriptorium repository, i.e. data from the National Library of the Czech Republic and most (but not all) Czech memory institutions.

5. Conclusions

The Manuscriptorium Digital Library has really managed to homogenize access to heterogeneous resources located in various places mostly in Europe. Thus, it offers a unique service over a large quantity of data and can serve not only users, but also other applications both created by memory institutions themselves, such as EUROPEANA, TEL, or CERL-MSS gateway, or operated as resource discovery services by known companies such as EBSCO or SUMMON. In this way the data aggregated through Manuscriptorium have a deep penetration in research communities. The statistics show that about 50% users are from outside of the Czech Republic.

Independently of all additional tools and services to which we should add, for example, also the Gaiji Bank[7] (database of non-standard/non-Unicode characters), the main Manuscriptorium mission remains the further aggregation of data consisting especially in recruitment of new content partners who would agree with parallel/secondary re-use of their data in the Manuscriptorium service. However, for the institutions that do not have how to go with their digital data in a more sophisticated manner onto Internet, Manuscriptorium is an opportunity how to do it quickly and with great usage impact, becoming thus the primary presentation site as it already is for an important number of small institutions.

[7] http://www.manuscriptorium.com/index.php?q=gaijibank