# ENRICH FINAL CONFERENCE PROCEEDINGS

National Library of Spain, Madrid

5-6 November 2009

# ENRICH
# FINAL CONFERENCE
# PROCEEDINGS

National Library of Spain, Madrid

5-6 November, 2009

## ORGANISING COMMITTEE

National Library of the Czech Republic, Prague
Biblioteca Nacional de España, Madrid
Cross Czech, a.s., Prague, Czech Republic
AiP Beroun, Ltd., Beroun, Czech Republic

## ENRICH PARTNERS

National Library of the Czech Republic, Prague
AiP Beroun, Ltd., Beroun, Czech Republic
Oxford University Computing Services, Oxford, United Kingdom
Centro per la comunicazione e l'integrazione dei media, Florence, Italy
SYSTRAN S.A., Paris, France
Institute of mathematics and informatics, Vilnius, Lithuania
Biblioteca Nacional de España, Madrid
Cross Czech, a.s., Prague, Czech Republic
Københavns Universitet – Nordisk Foskningsinstitut, Copenhagen, Denmark
Biblioteca Nazionale Centrale di Firenze, Florence, Italy
University Library Vilnius, Vilnius, Lithuania
University Library Wroclaw, Wroclaw, Poland
Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík, Iceland
Computer Science for the Humanities – Universität zu Köln, Cologne, Germany
St. Pölten Diocese Archive, St. Pölten, Austria
The National and University Library of Iceland, Reykjavík, Iceland
The Budapest University of Technology and Economics, Budapest, Hungary
Poznań Supercomputing and Networking Center, Poznań, Poland

# Index

# Presentation

ENRICH is a targeted project funded under the eContentPlus programme for the period 2007-2009. Its objective is to provide seamless access to distributed digital representations of old documentary heritage from various European cultural institutions in order to create a shared virtual research environment especially for study of manuscripts, but also incunabula, rare old printed books, and other historical documents. It builds on the Manuscriptorium Digital Library (http://www.manuscriptorium.eu) that has already managed to aggregate data from 46 collections from the Czech Republic and abroad.

The project groups together almost 85% manuscripts currently digitized in the national libraries in Europe, while its partners are also university libraries and other institutions such as foundations or special projects and initiatives.

The metadata records for the central database are collected preferably via the OAI protocol; they must contain links to images stored in remote image databanks. Necessary transformation routines are created and tuned for each partner. Specialized on-line tools are developed to enable Manuscriptorium schema compatible metadata structuring and output validation for those partners that have digital data with no presentation tools and will like to make them available.

The Manuscriptorium Digital Library is the largest digital manuscript library in Europe, it is accessible via The European Library TEL, and it constitutes a considerable digital manuscript segment for Europeana, the future European digital library.

The ENRICH consortium consists of 18 partners and the project is also supported by a number of other institutions among which there are many important content owners.

Our experience shows that dispersed information about manuscripts in Europe and in the world needs to be collected and their virtual representations offered in a homogeneous way to users. The scholars had to travel and they still have to in order to

reach the unique resources that are stored hidden in so many institutions in spite of their having close relationship as to their original provenance or contained ideas. The ENRICH-enhanced Manuscriptorium is a great chance for all of us to recreate formal collections and to offer powerful tools for study and research.

The final Conference of the project is to be held at the National Library of Spain and its main aim is to present the results of the ENRICH project and discuss the latest developments in the field of digital libraries with focus on the domain of manuscripts and early prints.

# Presentations

# SESSION 1

# THE ENRICH PROJECT AND WAYS OF COOPERATION

## THE ENRICH PROJECT AND WAYS OF COOPERATION

**Tomáš Psohlavec, Zdeněk Uhlíř, Adolf Knoll, Stanislav Psohlavec, Jakub Heller**
AiP Beroun, Ltd., National Library of the Czech Republic, Cross Czech A.S.

**Abstract:** Manuscriptorium (http://www.manuscriptorium.eu) is a resource provided by the National Library of the Czech Republic (http://www.nkp.cz) as a strategic leader and content coordinator as well as by the AIP Beroun Ltd. (http://www.aipeberoun.cz) as a technical provider and system administrator. Manuscriptorium became a major resource at the European level due to realization of the ENRICH project (http://enrich.manuscriptorium.com) funded under eContent+ programme. The project results are available for further usage and Manuscriptorium is open for cooperation with new partners providing various powerful presentation tools for their digital documents. The added value provided by Manuscriptorium in the area of digitized historical funds is well appreciated by both the cooperating partners and the end-users.
**Keywords:** *Manuscripts, Incunabula, Digital Library, Interoperability, shared repositories, On-line access.*

### INTRODUCTION

The aim of the ENRICH project is the creation of a base for the European digital library research environment for study of specific historical cultural heritage consisting of manuscripts, incunabula, early printed books, historical archival materials, etc. Practical validation of possibilities and definition of conditions for integration of existing but scattered electronic content under the existing Manuscriptorium digital library interface through the way of the metadata enrichment and coordination between heterogeneous metadata and data standards as well are the core objectives. The main innovation of ENRICH lies in a common easy-to-use interface which enables concentration of dispersed resources into a unique research environment and retrieval of data from distant servers. The project allows the users to search and access

documents which would otherwise be hardly accessible by providing access to almost all digitized manuscripts in Europe. During the initial session we will present the fundamental project principles, demonstrate aggregation results presenting selected partners documents/collections within the end-users interface and at the same moment we will demonstrate the Manuscriptorium end-users features with special focus on the newly available "Personalized Virtual Library" feature. The typical ways of cooperation will be also shortly mentioned using the real-life examples.

### Content Aggregation

The ENRICH groups together the richest owners of digitized manuscripts among national libraries in Europe; ENRICH partner libraries possess almost 85% currently digitized manuscripts in the national libraries in Europe, which is enhanced by substantial amount of data from university libraries and other types of institutions. The consortium will make available more than 5 076 000 of digitized pages by the end of November 2009.

The principle of integration is centralization of metadata (descriptive evidence records) within the Manuscriptorium digital library and distribution of data (other connected digital documents) among other resources within the virtual net environment. The project creates conditions that enable the partners (both the actual and those who will join us later) bringing together appropriate mass of digital content.

These conditions are open to approaches that may be applied by various institutions in the field of digitization of rare materials (national, university and other libraries and institutions holding historical documents). As we are aware that approaches to creating, maintaining and publishing of digital content vary in individual partners institutions, we prepared multiple different ways of cooperation and in addition we provide metadata conversion services (therefore no particular metadata format is required).

Manuscriptorium operates own harvester which enables automated transfer of documents originated in advanced digitization projects operating digital libraries equipped with the OAI-PMH interface. Various other methods of metadata transfers are supported.

### Manuscriptorium Platform

As stated above the ENRICH project builds upon the existing Manuscriptorium platform (http://www.manuscriptorium.com) adapted to needs of organizations holding repositories of manuscripts. Manuscriptorium is a resource provided on-line by the National Library of the Czech Republic (http://www.nkp.cz) as a coordinator as well as by the AiP Beroun Ltd. (http://www.aipeberoun.cz) as a technical provider and system administrator. The service is provided on a routine basis since 2003.

### Common end-user features

The partners digital manuscripts, early printed books and other historical documents benefit from the aggregation itself as these materials become highly accessible along with all related documents from various other aggregated funds. Similarly for the end-users the accessibility of documents is significantly increased.

Moreover the Manuscriptorium platform provides a set of powerful end-user features that make the digital content to be better usable.

A searchable Open catalogue of historical documents is an important part of the service along with the Digital Library that contains all digital documents aggregated so far. These end-users interfaces were significantly improved during the ENRICH project.

The search engine within the Open catalogue reflects specific user demands in the area of historical funds and makes the research process easy and efficient by provision of different search methods based on the types of users and their different demands.

Also the Digital Library interface is designed to provide intuitive tools for browsing digital documents and enables seamless incorporation of the documents from dispersed resources into a single presentation interface. The newly released version of the interface ensures even more comfortable navigation and image manipulation while again working with standard browser features (no additional plug-ins are required).

### Advanced end-user features: personalized digital library

Important part of works during ENRICH project were dedicated to **user personalization** in digital libraries. The activities concerning user personalization issues were initialized by all partners with the discussion about the actions which were undertaken to gather requirements for the creation of personalized virtual digital libraries. The conclusion was that in order to collect needed information partners prepared a set of questions in form of a survey for the end-users. The survey helped significantly to correctly recognize needs of interested users and collect opinions about functions needed.

As requested by the users the final development focused on implementing the possibility to subdivide the contents of Manuscriptorium into **thematic collections**. To satisfy the needs of all Manuscriptorium end-users, thematic collections were created and are maintained by authorized experts. Furthermore, end-users are able to construct their own individual collections and virtual documents by the means of newly developed tools – this creates the opportunity to build individual user virtual libraries according to their personal needs.

The newly available tools also allow to decompose the digitized documents into necessary chunks/analytical digital objects and recompose them in new **virtual documents** following special teaching or learning goals, e.g. showing all illuminations from one scriptorium in a virtual document in spite of the fact that they are from various originals owned by different institutions in different countries.

Such newly created content can be shared among users via Manuscriptorium interface, therefore we expect these features will be especially used to ease study, teaching and research tasks and also the usage of aggregated digital content will be increased accordingly.

### Benefits for cooperating partners

Apart from that Manuscriptorium provides partners documents with a set of powerful presentation tools there are other specific issues that increase the benefits of cooperation within ENRICH/Manuscriptorium project.

### Links to local digital libraries

Manuscriptorium links its users to the partners local digital library. The links of course can target various locations including the alternate local copy of digital documents. So the user can decide where to browse a document and also the local digital library can provide additional specific services.

### Inclusion of partners into other important EU projects

There are various front-end interfaces are implemented into the Manuscriptorium system including the OAI-PMH and Z39.50 based interfaces. Therefore inclusion of partners resources into selected portals is automatically achieved. For instance the Manuscriptorium DL is harvested via OAI-PMH by the TEL portal, i.e. any ENRICH full or associated partners document contributed to Manuscriptorium automatically enriches the European Digital Library.

### Tools supporting production of new digital documents

For starting or smaller scale digitization projects it is possible to use Manuscriptorium's dedicated tools to create digital documents metadata and transfer the documents to Manuscriptorium. These tools are:

- M-Tool: an on-line application which enables to create the description of the document and also to generate structural metadata (necessary information about structure of the original document and also information describing location of

related data (e.g. image) files); the output produced by the application uses the ENRICH TEI P5 schema which ensures high quality and longevity of the produced digital documents,

- M-Can: an on-line application dedicated to those who want to import the TEI P5 based documents directly into the Manuscriptorium environment; the application enables both to check correctness of documents and subsequent transfer for import.

## Conversion services

The Manuscriptorium internal environments is based on TEI P5 ENRICH schema. This TEI P5 fully compatible schema provides a complete suite of encoding possibilities, covering not simply the cataloguing and description of manuscripts or early printed books, but also the encoding of a digital edition in which metadata, digital image, transcribed text, edited text, and editorial annotation are all integrated in a standard framework.

All ENRICH partners contribute to the Manuscriptorium either directly using the TEI P5 ENRICH schema or indirectly by means of transformation process as the project is highly flexible regarding acceptance of metadata of various formats.

For all non-TEI sources (partners using MARC based formats, MODS, METS etc) a special connector performing conversion to TEI P5 ENRICH schema is prepared within Manuscriptorium. Most often these connectors are prepared individually in cooperation with each partner respecting various different approaches to metadata creation.

## Conclusion

All the tools and services described are available via http://www.manuscriptorium.com, in the time of publishing of this article some of the newly developed features are available as beta versions and will be subsequently released into a full service.

## References

All the tools and services described are available via http://www.manuscriptorium.com

Knoll, A., Mayer, T., Psohlavec, S., Vomlel, J. *Digitization of Rare Library Materials. Storage of and Access to Data: The Solution for the Compouhd Document, Manuscripts and Old Printed Books* [CD-ROM]. Praha: Národní knihovna České republiky, 1997.

Uhlíř, Zdeněk. "Standard MASTER: katalogizace rukopisů v XML". In: *Národní knihovna: knihovnická revue*. 13, 2002, Nr. 2, pp. 84-101. ISSN 1214-0678

_____. "Projekt 'MASTER' a problematika elektronického zpracování středověkých rukopisů". In: *Ikaros* [online]. 1999, č. 8. ISSN 1212-5075

_____. "Manuscriptorium na cestě k evropské digitální knihovně". In: Knihovny *současnosti 2007*. Brno: Sdružení knihoven ČR, 2007, pp. 136-144. ISBN 978-80-86249-44-5.

Chodorow, Stanley. "The Medieval Future of Intellectual Culture: Scholars and Librarians in the age of Elektron". In: *ARL: A Bimonthly Newsletter of Research Library Issues and Actions*. Issue 189, December 1996.

Giesecke, Michael. *Der Buchdruck in der frühen Neuzeit: Eine historische Fallstudie über die Durchsetzung neuer Informations- und Kommunikationstechnologien*. Frankfurt am Main: Suhrkamp, 1998. 957 pp. ISBN 3-518-28957-8;

O´Donnel, James J. *Avatars of the Word: From Papyrus to Cyberspace*. Cambridge, Mass.-London: Harvard University Press, 2000. 210 pp. ISBN 0-674-00194-X.

*Functional Requirements for Bibliographic Records: Final Report*. München: K.G.Saur, 1998. 136 pp. ISBN 3-598-11382-X

Uhlíř, Zdeněk. *Teorie a metodologie elektronicko-digitálního zpracování rukopisů a hybridní knihovna*. [The theory and methodology of electronic-digital processing of manuscripts and the hybrid library.] Praha: Národní knihovna České republiky, 2002. 324 pp. ISBN 80-7050-410-2.

Bryant, John. *The Fluid Text: a Theory of Revision and Editing for Book and Screen*. Ann Arbor: University of Michigan Press, 2002. 198 pp. ISBN 0472068156.

Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C. "Grid Information Services for Distributed Resource Sharing". In: *10th IEEE International Symposium on High Performance Distributed Computing*, pp. 181-184. IEEE Press, New York (2001)

Foster, I., Kesselman, C., Nick, J., Tuecke, S.: *The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report*, Global Grid Forum (2002)

# TEI P5 ENRICH SCHEME – METADATA STANDARD FOR THE DESCRIPTION OF MANUSCRIPTS

**Lou Burnard**
Oxford University Computing Services

**Abstract:** The ENRICH project illustrates how XML and web-based technologies promote greater access to the rich cultural heritage of European institutions, without compromising their inherent complexity or diversity. It builds upon many years of expertise in the development of technical and operational standards for metadata and encoding. Its use of TEI XML ensures that its outputs remain interoperable with new and developing systems worldwide.
**Keywords:** *Standardisation; XML; TEI; cultural heritage; manuscript cataloguing*

Concluding an excellent article[1] about the evolution of the TEI proposals for the description of manuscript materials, Matthew Driscoll makes the following observation:

> Attending the early meetings of MASTER and TEI-MMSS was, for the present writer at least but doubtless for others too, a bit like when as a youth one first has dinner at someone else's house and discovers that not everyone does everything in exactly the same way. It could be a small detail, such as how the table is set or the napkins folded, but it could also be something fairly major, like the order and composition of the courses: although pretty much everybody has their pudding last, some people eat their salad before, others with, and still others after the main course (but before the pudding,

[1] M. J. Driscoll. P5-MS: A general purpose tagset for manuscript description (Digital Medievalist 2.1, 2006) http://www.digitalmedievalist.org/journal/2.1/driscoll/

> naturally) –and then of course there are those who don't eat salad at all. We found at these early meetings that while there is quite clearly a single tradition for the description of (western) manuscripts, one with its roots in antiquity, there is also a great deal of variation within that tradition, and the majority of us are brought up and remain within one regional variety. An encoding standard for the description of manuscripts –or, for that matter, meals– needs to be flexible enough to accommodate this variation, while remaining true to the underlying tradition...

Unlike books, manuscripts are unique objects, often of great cultural value, which are typically catalogued locally by the many different institutions holding them. Great institutions are able to produce richly detailed, highly scholarly descriptions, while smaller or less well-resourced institutions cannot hope to do so. But with the widespread increase in the practice of digitization of such primary sources, there is increasing pressure to make their cataloguing uniform so as to facilitate cross-site searching. The World Wide Web makes it comparatively easy to share representations of the manuscripts held in all the collections of the world, but differing cataloguing practices, different views of what is essential, and different levels of resource, all make it difficult to share any but the most basic of such materials.

The ENRICH project is the latest of several initiatives which have addressed these difficulties. Its goals were:

1. to create seamless access to distributed information about manuscripts and incunables in Europe;
2. to connect existing digital libraries and to facilitate creation of new ones;
3. to enhance the existing Manuscriptorium system as a vehicle for providing such access to partners in their own languages using their own virtual interface;
4. to define and deploy a standardised metadata scheme based on the recommendations on the Text Encoding Initiative (TEI).

A major motivation for choosing a standardised metadata scheme and in particular one with the rather unusual degree of flexibility and modularity which characterizes the TEI scheme, was the support that such a choice would give to the other defined objectives of the project.

What strategies might one take faced with widely divergent cataloguing practices? One (typified by Dublin Core) might be to propose a kind of "lowest common denominator" of required categories for standardization. All users of the

standard would be required to provide data for each of a (relatively small) set of predefined and previously agreed concepts. Another (typified by RDF) would be to deploy a common language in which all of the concepts in each scheme under consideration can be re-expressed, in such a way as to identify automatically meaningful commonalities amongst them. The TEI scheme falls somewhere between these two extremes, in that it provides a large number (about 400) of predefined conceptual definitions, and facilitates selection from amongst them to form a particular "customization schema" or application profile. The hope is that the set of concepts identified by the TEI will be more or less coextensive with the set of unique concepts identified across each of the candidate schemas to be integrated; integration then becomes simply a way of mapping (for example) the terminology deployed by each candidate scheme into the normative terminology used by the TEI for the same concept. Surprisingly perhaps this rather simple-minded approach turns out to be a relatively effective one, at least in the problem domains where the TEI has been deployed hitherto.

Whereas previously manuscript catalogues primarily existed to benefit local users of a given collection, the advent of digitized versions of such resources, and consequently of digital descriptions of them, has brought some new opportunities. Manuscript descriptions stored in an online database become shareable, and searchable in new ways. They can be re-embedded in other kinds of publication along with much more discursive text to produce a more modern digital catalogue raisonné. They can form the metadata component of a complete digital surrogate (electronic edition) integrating image, transcription, and metadata; the TEI (as the standard of choice for such digital editions) is particularly useful in this context. And finally, we may want to deploy a wide range of software tools to search, count, and analyse a large number of our digitized descriptions, facilitating what has been termed a new 'quantitative codicology'.

The work done in defining a schema for the ENRICH project we believe helps achieve all of these goals. We have defined and implemented an expressive and reasonably complete conceptual model for the problem area, thus facilitating the lossless conversion of existing data, the creation of completely new data, and the integration of existing data from many different sources. We may note parenthetically that in basing this system firmly on open formats and open technologies we can reasonably have confidence that its long term maintenance and development can be sustained by its community of users.

In the ENRICH model, each manuscript description describes a particular object (no direct provision is made for no longer existent objects, nor for classes of objects).

Each description is organized using the same possible set of components:

- an identification for the object itself: its current and former shelf marks, nicknames etc.;
- descriptions of the 'intellectual content' of the object, using standard bibliographic practice where applicable;
- details of the object's physical makeup and composition: the carrier medium, the writing methods identified, the binding, and many more aspects including illustration and palaeography;
- records concerning the object's history, its provenance, ownership, curation etc.

Under these four broad headings, a very large number of discrete categories of information are identified by the model, each of which can be explicitly indicated by using the appropriate XML element. However, descriptions may also be informal or unstructured, in which case such elements will not be present. For ENRICH, few elements are mandatory.

In defining the ENRICH specification, care was taken to maintain compatibility with the full TEI P5 standard. A document prepared in conformance to the ENRICH schema is therefore also a TEI-conformant document, and can be used by any TEI-aware software. Furthermore, because of its use of the TEI, the ENRICH specification provides a complete suite of encoding possibilities, covering not simply the cataloguing and description of manuscripts or early printed books, but also the encoding of a digital edition in which metadata, digital image, transcribed text, edited text, and editorial annotation are all integrated in a standard framework.

For more information about the TEI please consult the website at http://www.tei-c.org.

# Towards deep searching in collections of old manuscripts by extracting semantic information

**Robert Kummer**
Historisch-Kulturwissenschaftliche Informationsverarbeitung
Universität zu Köln

**Abstract:** ENRICH can provide seamless access to distributed knowledge on manuscripts. For that, advanced information retrieval methods comprising complex linguistic, cross-language and simple semantic operations on metadata have been implemented. On this basis, this paper will discuss a simple use-case by introducing advanced semantic search facilities for metadata to enhance access to manuscripts. All ENRICH content partners agreed on providing knowledge on old manuscripts as TEI P5. And since the TEI provides – in addition to markup elements that are useful for describing manuscripts – means to record information about dates, people and places, the way for semantic processing has been cleared. A suite of software prototypes has been developed that implements a workflow to help prepare data which has been extracted from manuscripts for semantic browsing. The process of extracting information brought several obstacles to light that will be addressed. Possible solutions for these problems and relevant workings of Semantic Web research will be described.
**Keywords:** *cidoc crm, entity resolution, semantic web*

The task description of ENRICH announces the provision of semantic searches that shall introduce an "intelligent operator". But the notion of an "intelligent operator" still leaves a lot of room for interpretation and therefore this paper will elaborate a simple use case referring to current Semantic Web research for demonstration purposes. (W3C 2009) One way to approach this problem is to reflect on how applying semantic operators can enhance a users' research experience beyond that of a simple full-text search. A historian who is pursuing research on the life of a specific historical

person needs to acquire comprehensive knowledge about that person. Since names are notoriously spelled differently in old documents, a full-text approach will probably not be successful. Therefore, a system should be described that strives to provide semantic operators that are able to extract relevant bits of information from electronic manuscript descriptions.

Current Semantic Web research elaborated basic concepts and tools for information integration. In order to craft software components that implement the mentioned use-case these developments should be exploited. In this regard, different aspects of Semantic Web research turned out to be useful. One of the most fundamental concepts in this area is the notion of a Uniform Resource Identifier (URI) that provides a way to globally and unambiguously identify arbitrary material and immaterial things in the world. Furthermore, concepts like semantic markup and semantic triple stores have been exploited to facilitate semantic searches on ENRICH metadata. (Aduna 2009) To make use of these tools, certain information needs to be extracted from the ENRICH manuscript information.

A large amount of information in the humanities is derived from textual material. But even if texts have a clear structure and follow certain strains of arguments, from the perspective of automatic information processing they appear to be unstructured. In the context of ENRICH, all content providers agreed on providing information about old manuscripts as TEI P5 that comes with a certain predefined structure. (TEI 2009) First experiments showed that information about people, places and bibliographic entities could be extracted with reasonable effort. To support semantic searches that emancipate from simple field based evaluation strategies, the extracted information has been mapped to a common structured vocabulary, the CIDOC CRM. (Dörr 2003) We have decided for the CRM because it provides the needed structural elements to establish semantic interoperability in the cultural heritage area. The respective parts of TEI that deal more exhaustive with names, dates, people and places have been mapped to the CRM. (Eide & Emil-Ore 2006)

The notion of Linked Data has become quite popular in the area of Semantic Web research, aiming at explicitly linking related information to achieve better knowledge discovery. (Christian Bizer et al. 2009) In this context, one area of problems has been identified that inhibits proper semantic processing of knowledge called “object matching” or “entity resolution”. Historians for example are used to find references to historical people to be treated extremely inconsistent in old sources. Although, resolving these references is part of their day-to-day work, this task is laborious and extremely cost-intensive. (Eide 2008) Consequently, names that have been extracted from TEI documents do appear notoriously different although they are referring to the

same person. Resolving these references automatically could lead to unintentional results because there is no authority that is accountable for each matching decision. A semi-automatic approach seems to be the most viable approach. Therefore, the demonstrator provides an environment that helps with resolving the extracted name information. It makes use of simple data mining techniques to fuel a recommender engine. (Elmagarmid et al. 2007)

The performance of a co-reference recommender can be improved by exploiting information that has already been structured in a certain way. Authority control for example has been traditionally cultivated in library and information science where it is an integral part of bibliographic control. (Sieglerschmidt 2007) Authority lists help disambiguating items that share the same heading, and collocating material that belongs together but appears to be different. Thus, authority lists inherently document information about the aforementioned co-references. However, while traditional libraries have been good at curating these files, no human being will be in the position to fulfill this task on a larger scale with growing amounts of digitally enriched material. In the area of Semantic Web research, one developing standard for organizing knowledge stands out: SKOS intends to provide a more straightforward approach to publish multilingual structured vocabularies. (Isaac & Summers 2008) Initiatives like "museumsvokabular" (Stefan Rohde-Enslin 2006) publish their vocabularies as SKOS. This should be exploited in the course of work on information integration.

A number of functional requirements have been collected so far that project a future system to support semantic operators in the scope of ENRICH. Demonstrating the thoughts that have been elaborated so far, various software components have been developed that support a continuous workflow, beginning with information extraction and ending with visualization of the results. The paper will describe the implementation decision of each component. Additionally, a short excurse concerning artificial intelligence research will reflect on how an intelligent agent could be used to better support ENRICH users to fulfill their research needs. (Norvig & Russell 2003) These agents have been discussed as having knowledge about their users and therefore can independently perform certain research actions like informing a user about new documents that are related to a certain research topic.

## References

Aduna, 2009. openRDF.org: Home. *openrdf.org: Home*. Available at: http://openrdf.org/ [Accessed January 5, 2009].

Christian Bizer, Tim Berners-Lee & Tom Heath, 2009. Linked Data - The Story So Far. In *International Journal on Semantic Web & Information Systems,* Vol. 5, Issue 3, Pages 1-22, 2009.

Dörr, M., 2003. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag*, 24(3), 75-92.

Eide, Ø., 2008. What is co-reference? Available at: http://cidoc.mediahost.org/co_reference_wg%28en%29%28E1%29.xml [Accessed September 20, 2009].

Eide, Ø. & Emil-Ore, C., 2006. TEI, CIDOC-CRM and a Possible Interface Between the Two. In *Digital Humanities*. pp. 62-4.

Elmagarmid, A., Ipeirotis, P. & Verykios, V., 2007. Duplicate Record Detection: A Survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1), 1-16.

Isaac, A. & Summers, E., 2008. *SKOS Simple Knowledge Organization System Primer. Available at: http://www.w3.org/TR/skos-primer/ [Accessed November 25, 2008].*

Norvig, P. & Russell, S., 2003. *Artificial Intelligence: A Modern Approach* 2nd ed., Prentice Hall International.

Sieglerschmidt, J., 2007. Knowledge organization and multilingual vocabularies. Vortrag auf der Jahrestagung "Managing the global diversity of cultural information" des Comite International pour la Documentation (CIDOC) in Wien 20.-22. August 2007. Available at: http://opus.bsz-bw.de/swop/volltexte/2008/280/ [Accessed February 5, 2009].

Stefan Rohde-Enslin, 2006. museumsvokabular.de. Available at: http://museum.zib.de/museumsvokabular/ [Accessed September 20, 2009].

TEI, 2009. TEI: Text Encoding Initiative. Available at: http://www.tei-c.org/index.xml [Accessed September 21, 2009].

W3C, 2009. W3C Semantic Web Activity. Available at: http://www.w3.org/2001/sw/ [Accessed May 29, 2009].

# The role of selective metadata harvesting in the virtual integration of distributed digital resources

**Cezary Mazurek, Marcin Mielnicki, Tomasz Parkola, Marcin Werla**
Poznań Supercomputing and Networking Center, Poland.

Abstract: This paper presents the idea, role and benefits of selective harvesting extension of the OAI-PMH protocol, developed and applied in Polish digital libraries in frame of the European project named ENRICH (ECP-2006-DILI-510049), funded under the eContentPlus programme. Integration of scattered cultural heritage resources by means of the OAI-PMH protocol was one of the main objectives of the ENRICH project. Several digital libraries in Poland provide access to various digital documents including interesting for the ENRICH project cultural heritage documents. Unfortunately, not all digital libraries divide their content in such a way, that there is one or more specific collections of documents corresponding particularly to cultural heritage documents. To overcome this problem, Poznań Supercomputing and Networking Center, being one of the technical partners in the ENRICH project, chosen the approach to prepare a new extension to the OAI-PMH protocol. The extension allows for harvesting resources based on a search query specified in the Contextual Query Language. The solution is fully conformant with the OAI-PMH protocol, therefore does not influence unaware OAI-PMH harvesters. It also significantly decreases amount of transferred data between OAI-PMH data provider and OAI-PMH harvester. Furthermore, the OAI-PMH selective harvesting extension is applied to the Polish national aggregator – Digital Libraries Federation (http://fbc.pionier.net.pl/), which enables extended selective harvesting at the national level.

## 1. Introduction

ENRICH project is a targeted project, funded under the European eContentPlus programme. One of the basic aims for the ENRICH project is integration of existing but scattered digital cultural heritage resources such as manuscripts, incunabula, early printed books or archival papers. Manuscriptorium digital library is the integrating

portal for the project, initially developed by the National Library of the Czech Republic and AiP Beroun in scope of Memoria programme [1]. The preferred way for the integration is communication over the OAI-PMH protocol [2]. In case of content partners without ability to communicate over the OAI-PMH protocol, personalised integrating software tools are prepared.

In Poland, majority of digital libraries are based on dLibra system (http://dlibra.psnc.pl/), developed by Poznań Supercomputing and Networking Center (PSNC). dLibra system is fully compatible with the OAI-PMH protocol specification, therefore most of Polish digital resources can be harvested by external services, such as Manuscriptorium digital library. While digital libraries preserve documents of various types, the ENRICH project is focused on the integration of cultural heritage documents only, therefore selective harvesting had to be applied to gather only necessary documents. OAI-PMH protocol specification defines two types of selective harvesting criteria – date and set membership. The first one allows specifying harvesting criteria using date of creation, modification or deletion of the metadata record. Because the ENRICH project is gathering certain types of documents, this criteria is not useful. The second one allows for harvesting one of the predefined sets of digital objects and could be used for harvesting documents for the ENRICH project under one condition – there has to be a predefined set (or sets) of documents corresponding the cultural heritage in the harvested digital library. Unfortunately, not all digital libraries maintain sets of documents dedicated to cultural heritage, so it is not possible to gather necessary metadata in a simple and straightforward way. This problem can be solved by either fine-tuning selective harvesting on the content provider side or performing internal processing of gathered metadata on the integrating portal side. Because the first solution is more general, as it allows various harvesting projects to utilize this functionality, it was decided to introduce an OAI-PMH extension for selective harvesting functionality. The extension is based on the idea of a dynamic set which is defined ad-hoc by specification of the dynamic set membership criteria [3].

## 2. Dynamic set and selective harvesting

Dynamic set is a set of items, which is not defined in the digital library prior to the harvesting. The dynamic set is defined by specification of the membership criteria for the items. The criteria are passed from the harvester to the data provider, so the data provider is able to dynamically prepare the set and return all the items matching specified criteria. In case of vertical/thematic harvesters (such as Manuscriptorium system) the use of the OAI-PMH extension based on dynamic sets allows for harvesting metadata only in the scope of interest and decreases the

number of records transferred from the repository to the harvester. It is also very important that the harvester does not have to be aware of the selective harvesting extension and does not require any modifications because of the compliance with the OAI-PMH specification.

A dynamic set is a set for which a criteria for set membership is defined by the harvester. The only place where the criteria can be placed without adding additional parameters to the request is the set specification. So the OAI request can look like this:

*verb=ListRecords&metadataPrefix=oai_dc&set=SomeSet:EncodedCriteria*

which means that returned items should be from the *SomeSet* and additionally the items should match the *EncodedCriteria*.

To avoid a situation, where *SomeSet* accidentally would have a predefined subset with the specification that perfectly matches the *EncodedCriteria*, a special reserved word (e.g. "criteria"" in our case) could be used for a dynamic (sub)set specification. In such a case:

- for the query *&set=SomeSet:SomeSubset* - all items from *SomeSet:SomeSubset* will be returned,
- for the query *&set=SomeSet:criteria:SomeSubset* - all items from *SomeSet* matching the criteria *SomeSubset* will be returned,
- for the query *&set=criteria:SomeSet* - all items from the entire digital library matching the criteria *SomeSet* will be returned.

The criteria are encoded in the Contextual Query Language (CQL) [4]. CQL is a query language designed for various information retrieval systems. Its syntax is intended to be intuitive and readable and writable for humans. To conform to the restrictions of the OAI set specification, the CQL-based dynamic OAI set specifications should be URL-encoded (e.g.: dc.creator%3D%22Albert%20Einstein%22).

The proposed approach could not be strictly compliant with the current OAI-PMH protocol specification because of the nature of dynamic sets, but there is a solution to overcome this problem. There are two compliance problems. The first one is that the repository should list all its sets in the response to the *ListSets* request. The second problem is that the OAI-PMH specification requires that if a given item belongs to a set, then the set specification should be listed in this item metadata header. The solution is to replace dynamic set with the listing of one additional criteria subset for each set. Additionally if a harvest is done with a particular dynamic set, then this set can be listed in the items header.

Both problems described above should not cause any problems for a harvester that does not support dynamic sets. Dynamic sets may not be visible for this harvester or may look like empty sets. Therefore any OAI-PMH repository extended with the dynamic sets should be still OAI-PMH compatible for all protocol validators.

## 3. Conclusions

Selective harvesting extension for the OAI-PMH protocol enables various metadata aggregators to perform harvesting data providers based on search criteria which are applied to the metadata. As a result, returned records of metadata are only those matching given criteria.

This functionality has been successfully tested and is currently used in several digital libraries in Poland for the needs of various projects including European projects such as ENRICH (http://enrich.manuscriptorium.com/) or CACAO (http://www.cacaoproject.eu/).

Additionally, the same OAI-PMH extension has been applied to the Polish national aggregator –Digital Libraries Federation developed and maintained by PSNC– in order to allow various project such as Europeana, DRIVER or NDLTD to harvest metadata (selectively or not) at the national level [5].

Further works will focus on simplification of the OAI-PMH selective harvesting extension by adding possibility to alias search criteria in the data provider's internal configuration with simple and straightforward set specification. This will enable data providers to predefine dynamic sets and alias them with simple, human-readable specifications.

## References


Knoll, A., "Digital Access to Old Manuscripts". In *Linguistica Computazionale,* Digital Technology and Philological Disciplines, 277 – 286, 2004.

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. *The Open Archives Initiative Protocol for Metadata Harvesting*, 2002.

http://www.openarchives.org/OAI/openarchivesprotocol.html.

Mazurek, C., Werla, M. "Extending OAI-PMH protocol with dynamic sets definitions using CQL language". *Conference proceedings of "IADIS Information Systems"*, Algarve, Portugal, 2008.

Contextual Query Language specification, 2008.
http://www.loc.gov/standards/sru/specs/cql.html

Lewandowska, A., Mazurek, C., Werla, M., "Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland". *Research and Advanced Technology for Digital Libraries*, 12th European Conference, ECDL 2008, Aarhus, Denmark, 2008.

# Creating digital editions from medieval manuscripts or early prints: experiences of the National Library of Spain in ENRICH Project

**Bárbara Muñoz de Solano**
Biblioteca Nacional de España

Abstract: For end users of digital libraries it is not important the source of knowledge, but to get access to the information they want, and to be able to use significant materials from cultures around the world, including manuscripts, maps, books, musical scores, prints, photographs, architectural drawings, and other important cultural materials.
The purpose of this presentation is:

– To explain the contribution of the National Library of Spain on international projects related to the dissemination of ancient documents in digital format.
– To show strengths and limitations of *DigiTool* to participate in projects of international nature.

## Introduction

Since the end of the last century, European countries have invested in the digitalization of cultural collections, involving thousand of cultural institutions and private organizations such as archives, libraries, museums and others. Nevertheless, the fragile nature of ancient originals has limited access to these rich documentary sources while interest in the use of these manuscripts is increasing. The possibility of providing access through the use of digital copies is an attractive answer to the need to balance preservation and access. In this context the National Library of Spain considers that Biblioteca Digital Hispánica is able to play a double role in the recent "digital culture":

1. On one side, it can be an important vehicle for Spanish culture heritage diffusion at any time and any place;

2. On the other hand, Biblioteca Digital Hispánica can also be the mechanism of the National Library of Spain in order to participate in international projects related to the digitization of historical bibliographic materials:
   - Europeana: The European digital library, museum and archive – is a 2-year project. The intention is that by 2010 the Europeana portal will give everybody direct access to well over 6 million digital sounds, pictures, books, archival records and films. The digital content will be selected from that which is already digitised and available in Europe's institutions.
   - Enrich Project: Enrich project is focused on providing full access to distributed information about manuscripts and old printed books in Europe. The main objective is to create a virtual site especially for the study of manuscripts, but also incunabula, rare books, and other historical documents.

Although there are so many manuscript sites on the web[1] and a few online resources for palaeography[2] the ENRICH project is very much geared towards producing pragmatic and ready-to-use results. The project covers not just simple the cataloguing and description of manuscripts or early printed books, but also the encoding of a digital edition in which metadata, digital image, transcribed text, edited text, and editorial annotation are all integrated in a standard framework.

## Spanish National Library Contribution to ENRICH Project

To contribute to Enrich project and improve the presence of manuscripts and old printed books around the world, the National Library of Spain has created the following digital collections:

**Incunabula:** Adobe PDF file document
Technical information

[1] Manuscripta Medievalia, http://www.manuscripta-mediaevalia.de/ ; Bestiaire Mediavale: http://expositions.bnf.fr/bestiaire/index.htm ; Gastronomie medievale: http://expositions.bnf.fr/gastro/index.htm ; Digital scriptorum: http://www.scriptorium.columbia.edu/ ; Illuminating the Law: http://www.fitzmuseum.cam.ac.uk/gallery/law/index.html ; Medieval manuscripts: http://libwww.library.phila.gov/medievalman/ ; Medieval Manuscripts in the National Library of Medicine: http://www.nlm.nih.gov/hmd/medieval/medievalhome.html

[2] English Handwriting: http://www.english.cam.ac.uk/ceres/ehoc/ ; Medieval Writing Paleo Anglo-Norman: http://www.medievalwriting.50megs.com/ ; Paleography tutorial: http://paleo.anglo-norman.org/ ; Scottish Handwriting (16th-18th centuries) http://www.scottishhandwriting.com/

Using the latest scanners and image enhancement software the National Library of Spain converted 35 mm microfilm of incunabula documents into digital image files. The microfilm conversions were done at a resolution of 300dpi to produce the best possible images and trying to maintain high accuracy rate during the conversion process. Microfilms of incunabula were scanned, processed for full-text retrieval[3] and converted to Adobe PDF format after extensive quality inspection.

**Manuscripts:** JPG format as complex document.
Technical information
The digital collection of manuscripts integrates a rich and exclusive selection of historical documents held in by the National Library of Spain. It hardly needs to be said that these resources were selected by a group of wise specialists in many aspects of Sciences and Culture, including Literature, Art, Law, Linguistic and History. The mission of this digital collection is to make remarkable resources (among others, Beato de Liébana, the Cantigas de Santa María by Alfonso X el Sabio, the Codex Madrid I & II by Leonardo da Vinci, De aetatibus mundi imagines by Francisco de Holanda) available and usable for future generations. The images were captured on a digital camera, and edited to create master copies. Copies were then made from the master files, digital embedded watermarks were created for them and the size of the images was reduced. The image sizes have been reduced from the original master files into size that can be seen using our Digital Library display. ID numbers were given to the images on JPG format, titles and descriptions. Descriptions of the images were then incorporated into the XML files containing the catalogue data relevant to each volume. The purpose of the table included below is to describe the main characteristics and dimensions of documents ingested in the National Digital Library of Spain:

| | Format | Bit Depth | Resolution |
|---|---|---|---|
| Incunabula | PDF | Black and White | 150ppp |
| Manuscripts | Tiff and JPG<br>Complex documents | Colour | 300ppp |

[3] Two OCR engines have been employed Abbyy Fine-Reader 9.0 and Omni page; but both OCR output were very low accuracy of 20%.

**Software Analysis**

DigiTool is a well-respected system designed to facilitate the management of digital documents. Some key benefits of the system are:

1. The use of international standards.
2. Different types of media files can be deposited (e.g. mp3, pdf, video)
3. OAI harvesting
4. Different authentications can be set up for different users
5. Copyright can be managed by manual assignment of access rights to the object.
6. The major advantage is that general users, not only researchers, can consult old documents and manuscripts of the National Library at any time, night or day.
7. The easy management of both digital objects and descriptive metadata.
8. Via the Deposit Module. The Deposit Module provides an interface and workflow which enables submission of objects and metadata by non-staff users.
9. Workflow stages can be configured "to some extent", so that a central library service can monitor self-submitted documents for quality control and copyright issues
10. The use of persistent identifier
11. Documents can be set to open or closed access

Nevertheless, Digitool is not a perfect system. There are some elements included in Digitool that are not operational yet and also pending issues that should be improved in order to enable the software more sustainable to manage huge collections of digital documents. Let me mention just a handful of items here that I find particularly compelling:

1. Error log file mechanically created after the ingest process should be clear. We have notice that it is particularly important that once the failure has properly been identified by the system it would have to be individually linked to the mistake.
2. Currently it is not possible to manage PREMIS metadata.
3. Since UTILITIES_Print History function does not work librarians can not make a list of documents.
4. Special characters are not allowed in file and folder names. Because of the complexity of METS records, DigiTool provide for librarians the alternative of DTL Naming Convection tool. Nevertheless, this tool can not be an option for the National Library of Spain, because special characters are not allowed (for instance *ñ*).

5. As digital library contents are not static it is very important that collection management module could have an easy migration and import process of our existing digital collections.
6. Nowadays Digitool uses a proprietary plug-in to display JP2 Documents. The system should offer the possibility of using an open source Jp2 plug-in (for example Lizartech)

## Conclusions: where do we go from here?

The following issues will be the focus of our efforts in the near future:

- Increase the volume of our digital collections
- Discover new ways to use Digitool to provide better library service. Improving access and usability by:
  – Learning from users
  – Adding a number of new services
  – Working on user personalization
  – Providing translations of collection descriptions.
- Capture and describe digital works using customized workflow processes
- Cooperate with editors in order to improve the number of contemporary e-books in our digital library.
- Preserve digital works for the long term.

# Case studies of digital libraries cooperating with Manuscriptorium

Session 3

## The National Library of Romania in the European Digital Library of Manuscripts ENRICH Project

Nicoleta Rahme, Mariana Radu, Luminita Gruia

National Library of Romania has the mission to preserve and to ensure access to the Romanian cultural heritage. Its activity is in accordance with the European framework regarding the process of Digitization of the Cultural Heritage, especially the written documentary patrimony.

National Library of Romania is partner in relevant international projects, like TELplus, ENRICH, REDISCOVER.

National Library of Romania Cultural values consist of:

a. Special Collections

- Foreign Books
  - 142 incunabula
  - 17.950 old books (16th – 18th century)
  - 4.961 rare books (19th – 20th century)
- Romanian Books
  - 2.250 old books (16th – 19th century)
  - 6.153 rare books (19th – 20th century)
  - old and modern manuscripts

b. Batthyaneum Collections

- over 65.000 bibliographic units, among which: 61.683 old and rare books (Romanian and foreign); 603 incunabula; 1.600 manuscripts (9th-18th century); archival documents; museal collections

- the fameous manuscripts:
  – Codex Aureus (9th century)
  – Codex Burgundus (15the century)

ENRICH is a targeted project funded under the eContentPlus programme. Its objective is to provide seamless access to distributed digital representations of old documentary heritage from various European cultural institutions in order to create a shared virtual research environment especially for study of manuscripts, but also incunabula, rare old printed books, and other historical documents. It is built on the Manuscriptorium platform (http://www.manuscriptorium.com) that has already managed to aggregate data from 46 collections from the Czech Republic and abroad.

National Library of Romania is involved in ENRICH project since may 2008.

The contribution of National Library of Romania to the ENRICH project consists of old Romanian books from the XVI - XVIII centuries, of outstanding cultural, historical and artistic value. Most of these treasures are religious works, but also law and history books, realized by representatives printers for South-Eastern European space. Also, 357 rare and valuable manuscripts from Batthyaneum collections, will be integrated until the end of 2009.

The documents selected to be integrated in Manuscriptorium portal were already scanned within a local project. The selection criteria were value, age, bindings or different adnotations (handmade) and the conservation status of the books.

At present, 109 documents from Special Collections (old romanian books) are accessible in digital format on *www.manuscriptorium.com* and other 123 documents from Batthyaneum collections (valuable manuscripts) are uploaded in *http://candidates.manuscriptorium.com*.

**Methodology of work**

The digital content is stored on the local server of the library (storage server, accessible via http protocol), while the metadata records, created with the tools offered by the Manuscriptorium portal, are uploaded in the central database of Manuscriptorium.

- **MTool** - bibliographical and technical description - 3 specialist librarians in old books and manuscripts have created the bibliographical descriptions - metadata = *.med* file
- ICT Department was in charge with:
  – image processing
  – images uploaded on NLR server

– technical metadata - numbering, links
– review
– uploading in ***candidates.manuscriptorium.com***

- XML record - Manuscriptorium builds on a robust xml schema the most important part of which is the European MASTER format for electronic description of manuscripts based on TEI
  – fields provided
  – brief description of the data to be entered
  – incorporate content into the elements of DTD Master +

The xml files are uploaded in http: //candidates.manuscriptorium.com. Records are verified by the National Library of Czech Republic and diacritics added. National Library of Romania makes the final revision and the record is uploaded in http://www.manuscriptorium.com.

**Slavonic Book of Liturgies (Liturghierul lui Macarie)**

**Macarie Slavonic Book of Liturgies *Ije Svjatyh Otta nasego arhiepiskopa Kesarie Kapadokinskaja Vasilia Velikago pooycenie k" preazvyteroy o boj'st'viean sloyjbea i o pricescenij***

Is the first book printed on the present-day territory of Romania, in 1508, during the reign of Radu the Great, by Macarie of Montenegro extraction. The watermarks of the paper –representing a balance with round or triangular pans in a circle, an anchor in a circle and a cardinal's hat– are proof that it was made in Italy (most probably in Venice). The language of the text is Middle Bulgarian.

**ENRICH/Manuscriptorium in Romania**

- presented in national and international conferences - expose cultural heritage
- articles and presentations dedicated to this project
- many libraries expressed their interest in participating in the project
- different libraries - to contribute with local written documentary patrimony
  – *Central University Library Bucharest*
  – *Romanian Academy Library Bucharest and Cluj*
  – *County Public Library Brasov*
  – *County Public Library Targu Mures*
  – *County Public Library Galati*

## Perspectives

At the end of the year 2009, all the 357 scanned manuscripts from Batthyaneum collections will be uploaded in the portal. Also, other manuscripts, historical archive documents and incunabula that were scanned in 2009 will be integrated in the portal.

**Manuscriptorium portal** is integrated in The European Library www.theeuropeanlibrary.com

## Conclusions

The ENRICH project groups together almost 85% currently digitized manuscripts in the national libraries in Europe. National Library of Romania owns valuable collections that include some of the most important prints and manuscripts from the South-Eastern European heritage, that is now accessible in digital format, through Manuscriptorium portal.

Thus, through **ENRICH/*Manuscriptorium***, digital content based on diversity is provided, and National Library of Romania valuable collections are a visible and accessible part of the european cultural heritage.

# *Handrit.org*: A Digital Library of Icelandic Manuscripts

**Matthew James Driscoll, Eric Andrew Haswell**
University of Copenhagen, Denmark

**Abstract:** This papers present the collaborative effort of three institutions in the establishment of a digital library of Icelandic manuscripts, *handrit.org*, based on the work of the ENRICH project. The institutions, all of which are partners in ENRICH, between them hold nearly 90% of the Icelandic manuscripts extant. *Handrit.org* was conceived as a central point of access for information about and analysis of the manuscripts in these three collections. The system, which is currently in beta development stage, is based wholly on the native XML database eXist, with PHP used for the website front end. TEI-conformant XML manuscript descriptions are produced according to the ENRICH schema. These provide information on the manuscripts' contents, physical structure, origin and subsequent history. Controlled vocabularies are used to regulate content, typically through fixed lists of attribute values defined in taxonomies or 'hard wired' into the schema. Extensive use is also made of authority files, for example for the names of persons, places and institutions, using the TEI elements <listPerson>, <listPlace> and <listOrg>, respectively. By combining various criteria a nuanced picture of Icelandic manuscript production and consumption over many centuries can be obtained.
**Keywords:** *XML, TEI, XQuery, PHP, XML databases, manuscript cataloguing*

## Introduction

The Arnamagnæan Manuscript Collection, recently inscribed on UNESCO's 'Memory of the World' Register, derives its name from the Icelandic scholar and antiquarian Árni Magnússon (1663-1730). The collection comprises nearly 3000 items, the earliest dating from the 12th century. Around three-quarters of these are Icelandic, the

remainder being chiefly Norwegian, Danish and Swedish, as well as some of continental provenance. Following repeated petitions from Iceland, until 1944 part of the Danish realm, roughly half the collection was transferred to Iceland, a process completed in 1997. The manuscripts in Iceland retain their original shelfmarks, and the collection is jointly administered by the Arnamagnæan Institute (*Den Arnamagnæanske Samling*) in Copenhagen and the Árni Magnússon Institute for Icelandic Studies (*Stofnun Árna Magnússonar í íslenskum fræðum*) in Reykjavík.

The manuscript collection of the National and University Library of Iceland (*Landsbókasafn Íslands-Háskólabókasafn*) in Reykjavík, established in 1818, comprises some 15,000 items, the bulk of them paper manuscripts from the 18th and 19th centuries. Between them, these three institutions hold nearly 90% of all the Icelandic manuscripts extant[1].

*Handrit.org* was conceived as a central point of access for information about and analysis of the manuscripts in these three collections. *Handrit.org* is currently in beta development stage. Stability and functionality is sub-optimal and work is ongoing in all aspects of the system's development. The main thrust of the work in building *handrit.org* has been twofold: 1) the development of the database system and the technical infrastructure underlying it; and 2) electronic cataloguing.

## The database

The system is based wholly on the native XML database eXist. It is the nature of document-centric XML data, such as manuscript descriptions, that they cannot be 'shredded' and fed into the table-based structure inherent to a relational database system.

Indexing is performed internally and automatically by the eXist database when a document is added or changed. Several different types of indices are supported. A basic structural index indexes the nodal structure, elements and attributes of a document and of the documents in a collection. Range indices provide a shortcut for the database to select nodes based on their typed values directly. A full text index is also available, as are several other types of indices.

## Web application

PHP is used to develop the website front end. It handles basic things such as: page construction from modular content, tracking user-state and determining which of the three interface languages –Danish, Icelandic or English– is to be used.

[1] Other significant collections of Icelandic manuscripts are found in the Royal Library in Copenhagen, the Royal Library in Stockholm (293 items), Uppsala University Library, the British Library and the Bodleian Library in Oxford.

To query the manuscript data, XQuery is used, a language with capabilities similar to those of SQL, but tailored to XML data and thus possessing of a completely different syntax. The XQuery engine is built into and is an integral part of the eXist database. The XML resulting from an XQuery is passed to an XSLT engine (in this case Saxon 9B) which prepares the content for output to the web.

The web interface is XHTML 1.0, with a considerable amount of JavaScript to enhance usability. Notable in this context is the implementation of some components of the Yahoo User Interface Library, a collection of JavaScript widgets.

The basic structure is a standard three-tier web database application, with PHP handling communication between the web client tier and the database tier. Data are received by PHP from the client in the form of request variables. PHP passes these to the database via a RESTful web service call.

## XML SCHEMA AND TEI CONFORMANCE

The work of cataloguing has involved either converting existing catalogue records –the National and University Library, for example, had several thousand records in MARC format– or producing new ones in XML, following the recommendations set out in the latest version of the TEI Guidelines, P5[2]. The schema used is in fact a narrow subset of P5, which was specifically developed by and for the ENRICH project[3]. It includes only those elements needed for the description and transcription of primary sources, as well as elements for linking these descriptions and transcriptions to digital images, where they exist.

A range of elements is employed to provide information on the manuscripts' contents, physical structure, origin and subsequent history. Controlled vocabularies are used to regulate content, typically through fixed lists of attribute values defined in taxonomies in the TEI header or 'hard wired' into the schema. One example of the former is the list of possible text-types available as values of the @class attribute on <msItem>. This list is based on collaborative work by Icelandic and Danish manuscript scholars and does not represent a 'standard' as such, though it might well become one. In other cases existing international standards are used, and the value lists built into the schema. When recording a person's gender, for example, the value of the @sex attribute on <person> may only be '0', '1', '2' or '9', in keeping with ISO standard 5218:1977, 'Representation of Human Sexes'; 1 and 2 indicate male and female respectively, while 9 indicates not applicable and 0 unknown.

[2] *Guidelines for Electronic Text Encoding and Interchange* (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html).

[3] The ENRICH schema: *A reference guide* is available at http://tei.oucs.ox.ac.uk/ENRICH/ODD/enrich.xml.

Extensive use is also made of authority files, e.g. for the names of persons, places and institutions, using the TEI elements <listPerson>, <listPlace> and <listOrg>, respectively. All proper names occurring in the individual manuscript descriptions are tagged using <name>, with a required @type attribute to indicate whether it is the name of a person, place or organisation/institution and a @key attribute which points to the relevant <person>, <place> or <org> element. Shown here is the <person> element for Jón Erlendsson, a 17th-century Icelandic clergyman who copied many manuscripts.

```
<person sex="1" role="scribe" xml:id="JonErl001">
    <persName xml:lang="is">
            <forename sort="1">Jón</forename>
            <surname sort="2">Erlendsson</surname>
    </persName>
    <birth notBefore="1600" notAfter="1610"/>
    <death when="1672-08"/>
    <residence>
            <placeName>
                    <settlement type="farm" key="#VilVil01"/>
            </placeName>
    </residence>
    <occupation key="#pr"/>
</person>
```

Jón's dates are given as empty elements, intended principally for search purposes. For display purposes, however, appropriate content can be generated from the attribute values; the date of death, for example, can appear as 'August 1672', 'august 1672' or 'ágúst 1672' depending on whether the interface language selected is English, Danish or Icelandic. Occupations are dealt with in a similar fashion; here the value of the @key attribute resolves to 'clergyman', 'præst' or 'prestur' as appropriate.

Jón's place of residence is similarly given as an empty <settlement> element, the @key attribute of which points to the relevant <place> element in <listPlace>:

```
<place xml:id="VilVil01">
    <placeName xml:lang="is">
            <settlement type="farm">Villingaholt</settlement>
            <region type="parish" key="#Villin01"/>
    </placeName>
```

```
    <location>
            <geo>63.883997 -20.750909</geo>
    </location>
</place>
```

Following the name of the settlement, further information is given as an empty <region> element, also with a @key which points to the relevant parish, defined in a separate <place> element, from which there is a pointer to the county, from there to the geographical region and so on. This strictly hierarchical structure ensures that information is only given once, preventing repetition and the possibility of conflicts. For each <place> element precise geographical co-ordinates are given in order to be able to locate the places on a map.

In this way it is possible to search for manuscripts written at a certain time, in a certain place and containing certain types of texts. By combining these criteria with others relating, for example, to the social status of the scribes and owners and, say, manuscript format, a nuanced picture of Icelandic manuscript production and consumption over many centuries can be obtained.

# Monasterium-net-
# A virtual archive for european charters

**Karl Heinz**
Diözesanarchiv St. Pölten, Austria

**Abstract:** Monasterium.Net is world-wide the biggest online database focused on medieval and Early Modern charters. At present 60 partners from 10 countries participate in the project and contribute more than 120.000 single charters on the World Wide Net. The charters are presented with a digital picture and a collection of various metadata. Beside the pure presentation on the internet Monasterium offers an opportunity to collaborate by the editing tool EditMOM. Registered users are entitled to create new metadata or make corrections. The quality of the new data input is guaranted by a moderation system. The editing tool is based on the CEI-standard (Charter Encoding Initiative) and allows a user friendly semantic tagging of the texts by the user community. The system offers completely new possibilities for researchers/historians and new teaching methods at universities on the field of diplomatical, paleographical and archival education.
**Keywords:** *charters, digitization, archives, Middle Ages, standards, collaborative tool, EditMOM, monasteries, education, network*

Austria and especially the federal province of Lower Austria (in the north-east of Austria) has a very high density of still existing abbeys of the "old orders", that still exist from the Middle Ages –mostely from the 11th century on– till today. Speaking about the "old orders" usually the Benedictine, Cistercian, Augustinian and the Premonstratensian orders were ment. These monasteries still keep their written inheritage in their archives.

Starting from their foundation every monastery was part of the europewide network of the different orders they belong. The charters are mirroring this medieval

connections in a big variety. These archive stocks deal not only with the history of the abbey itself, but cover social, economical and arthistorical belongings, too. So we can say, that they represent not only an important historical source but at the same time are a peace of identity not only for Austria, but for big parts of East-Central Europe, too. So there is a very strong relationship between medieval charters and cultural heritage. The question was, whether there is a way to position this fact in public consciousness.

In spite of the historical importance of these sources for local and regional and even for european history, accessibility was quite poor. In the year 2002 the idea arose to digitize the ecclesiastical charter stocks of the Lower Austrian monasteries all together about 20.000 pieces. The project´s name was MOM - the commonly used abbreviation for latin *monasterium* - and aimed to transfer the historical connections layd down in the medieval charters into modern networks of a digitized world and make historical sources available to everybody, who has access to the word wide web.

Due to the intensive international connections of the ecclesiastical institutions it was obvious, that the project could not stop at the borders of Lower Austria and even has to go beyond the Austrian borders. Till today (september 2008) more than 60 partner institutions participate in the project providing more than 120.000 charters via the Monasterium portal (www.monasterium.net). The partners are located in Germany (Bavaria), Czech Republic, Slovakia, Hungary, Slovenia, Croatia, Italy, Serbia, Austria and Switzerland. Among them are most of the national and state archives, provincial and municipial archives and the most important ecclesiastical archives.

The Monasterium portal offers two different possibilities of approach. One way ist mainly passive and limited on viewing the material. The charters are presented via a digital picture and a collection of various metadata. Due to a quite high resolution of 400 dpi (TIFF-Masters) the images have a higy zoome factor which maximize the legibility of the original. The metadata offer all aspects of diplomatical exploitation like date, summary, transcription, issuer, seal description, measures, language etc.

On the other hand Monastreium is also an interactive platform, offering a big variety of collaborative possibilities. After potential users had gone through the process of registration they can use the integrated online editing tool EditMOM in order to emend already existing data or to encrease the present data status. Data can be added in two ways. On a first level information like transcriptions, summaries, provenience, copy traditions, seal descriptions etc. can be filled in into the provided data fields. On a second level already existing texts can semanticaly be marked up additionally. Tagging is based on the XML standard of the CEI (Charters Encoding Initiative), which is based on the TEI standard, enriched by specific and necessary elements to describe medieval charters properly. The CEI standard (http://www.cei.uni-muenchen.de/) has

been developed by an international working group mainly at the Ludwig-Maximilians university in Munich (Georg Vogeler).

EditMOM is designed in a quite user friendly way, following the model of the commonly used text processing software solutions. Elements (tags) can be inserted menu-driven in a confortable way. Mark-up can be done in three different contexts. First the charter can be analysed in a formal-diplomatic way by tagging abbreviation, additions, deleted words or characters, corrections, damages, diacritical characters, mistakes, handshifts, highlighted text, paragraphs and so on. Secondly the editor can deal with the content by marking up person names, place names, geographical names, witnesses, dates and time periods, numbers, measurements, citations, alternative languages and so on. On a very specific level it is also possible to mark-up the formal structure of a charter by identifying the different constituting parts of a charter, like invocatio, arenga, narratio, dispositio, corroboratio etc.

One very important point is to secure the quality of the new data input, which is guaranted by a moderation system. If a user decides to collaborate in the system, he or she has to choose during the process of registration a moderator obligatorily, who is in charge of the finaly published content on the web.

A big advantage is that, due to the online status, collaborative action can take place when and wherever in the world without any restrictions of space and time with the only precondition of an internet connection. In most cases regarding the content of a charter it is not necessary any longer to use the originals. So the researchers need not travel to each single archive to see the originals. This spares costs and means on the other hand less work for the archivists, who now are enabled to invest this won time in other areas of activity. Another advantage of course is less pysical strain for the originals, what grants a longer durability of the charters.

Due to the international consistence of the partners and the data-stocks the Monasterium portal as well as the editing tool EditMOM is multilingual (german, italian, czech, slovak, hungarian, croatian, serbian, slovenian and english) so that everybody accessing from one partner country can navigate and work in a familiar surrounding.

Besides of the new possibilities for historical research of queries in the context of more than 120.000 charters from 10 countries, which are obvious, historical and archival education can be transformed and modernised by using the editing system. At universities EditMOM can be used as a supporting tool in seminars and tutorials dealing with paleographic and diplomatic matters.

Another result of the cooperation in order to create the virtual charter´s archive is the development of a compact network of partners, who form the Monasterum

consortium. In order to support know how exchange and to prepare the common targets, this assembly comes together in an interval of half a year.

In order to subsume the results and benefits of the Monasterium project, that have been achieved in the last seven years it can be said that Monasterium

- assures a comfortable, multilingual and free access for anybody interested in history (scientists, local researchers, students, teachers etc.)
- enables queries in more then 120.000 charters from 10 different countries (planned to be enlarged in the next 2-3 years up to 300.000 charters)
- as a virtual archive helps researchers to spare money and time and is a contibution to extend life duration of the original parchments
- makes efforts to develop commonly shared and accepted technical and scientific standards (CEI)
- gives the possibility of emendation and augmentation of the data with a user friendly collaborative online editing-tool (EditMOM)
- is a contribution to strengthen the common historical traditions in East-Central Europe while appreciating the regional plurality

# Heidelberg University Library-Partner of manuscriptorium/ENRICH

**Dr. Karin Zimmermann**
Heidelberg University Library, Germany

**Abstract:** Since September 2008 Heidelberg University Library contributes its 848 digitized German language manuscripts of the former Bibliotheca Palatina to Manuscriptorium and the ENRICH Project. The Bibliotheca Palatina is regarded as one of the most valuable collections of medieval and early modern manuscripts in the German language. The manuscripts are dating from the late ninth to the early 17th century. Its origins go back to 1386, the date of the founding of Heidelberg University. At the beginning of the 17th century it became the biggest and most famous library in Germany. During the Thirty Years' War it was taken as booty to the Vatican Library in Rome, where nearly all non-German manuscripts and all prints are still kept. In 1816 the German manuscripts were returned to Heidelberg University Library. After a first contact of members of Manuscriptorium and the library in December 2007 the IT-departments checked the technical standards required for the integration of the digitized manuscripts and the metadata into the database. A short phase of testing was soon followed by routine harvesting via OAI-interface.
**Keywords:** *data transmission, digitization, manuscripts, METS, MODS, Dublin Core*

## 1. Digitization of the German language Palatina manuscripts

I first met Zdeněk Uhlíř, the coordinator of "Manuscriptorium" at the National Library of the Czech Republic, after my talk at the LIBER conference in Berlin in December 2007, where I introduced the Heidelberg project of the digitization of the German language Palatina manuscripts. Zdeněk asked me if Heidelberg University Library would be willing to cooperate with "Manuscriptorium/ENRICH" and provide these projects with its digital facsimiles of the mentioned manuscripts of the Bibliotheca Palatina.

The Bibliotheca Palatina at Heidelberg is regarded as one of the most valuable collections of medieval and early modern manuscripts in the German language. It consists of 849 manuscripts dating from the late ninth to the early 17th century. Its origins go back to 1386, when the University of Heidelberg was founded by Elector Ruprecht I. During the following centuries the collegiate library of the Heidelberg Heiliggeistkirche (Holy Ghost Church) and the private book collection of the Palatinate Electors were incorporated into the growing University Library, until it became the biggest and most famous library in Germany. During the Thirty Years' War it was taken as booty to the Vatican Library in Rome, where today nearly all non-German manuscripts and all prints are still kept. In 1816 –after the Napoleonic wars– the German manuscripts were returned to Heidelberg, where they are preserved in the University Library.

Because of the fragile condition of some of these books they are no longer accessible to the public. Therefore we decided to digitize the whole collection to reduce the use of the originals to a minimum.

The aforesaid project "Digitization of the German language Palatina manuscripts" had been running at Heidelberg University Library since May 2006 and ended successfully in April 2009. It was supported by a foundation, the Manfred-Lautenschläger-Stiftung[1].

The digital photography was carried out in our digitization centre where we used two so-called "Graz book tables". This kind of book table permits non-contact, direct digitization. The book becomes accurately positioned with the aid of a laser beam, so that the camera is always at right angles to the manuscript and distortion is minimized. The pages are fixed one at a time with low pressure suction and the aperture angle of the book is reduced to a minimum.

Since 2008 our IT-department additionally developed a program –we called D-Work– to manage the workflow of digitization and internet-presentation of our manuscripts (and prints). On the one hand the program generates the presentations, but with its help we can also control the long-term archiving of scans and metadata. Furthermore, it automates and depicts every single step of the workflow, so we are always able to control how far each process of digitization and presentation of a manuscript or print has gone.

Within three years all German manuscripts of the Bibliotheca Palatina were digitized; in total about 270.000 pages and 6.500 miniatures. That means, on average

[1] Effinger, Maria ; Krenn, Margit ; Wolf, Thomas. "Der Vergangenheit eine Zukunft schaffen". *Die Digitalisierung der Bibliotheca Palatina in der Universitätsbibliothek Heidelberg*. In: B.I.T. online 11 (2008), Nr. 2, S. 157–166.

we released one digital facsimile of a manuscript each working day. Without the sponsorship and the third-party funds, only supported by the regular budget of the library, the project would have lasted over 20 years. But so since April 2009 all manuscripts are online and all pages with miniatures are indexed. The listing of the shelf-mark ordered digitized codices is presented on the web site of Heidelberg University Library (http://palatina-digital.uni-hd.de).

So far the short information about our digitization project that provided the basis for our collaboration with "Manuscriptorium/ENRICH".

## 2. Integration of the Heidelberg data in Manuscriptorium/ENRICH

Back to Heidelberg - after the Berlin conference -I immediately conferred with the director of our library, Dr. Probst. He made the basic decision to participate in the database "Manuscriptorium" with "free data" (we only contribute bibliographic metadata and links to our digital facsilmiles, so we prevent to mirror our images on the Prague server) to make sure that there is no restriction concerning the free use of our digitized manuscripts. He asked me to check the required technical standards and to make contact with our IT-department. They didn't see any problems concerning the integration of our data into the "Manuscriptorium" database. The only problem was, that at this time we couldn't offer our data in METS (Metadata Encoding & Transmission Standard). Because this would make the data transmission simple and easy to maintain our technicians preferred to wait with the integration until then (announced for the first quarter of 2008).

Nevertheless as a first effective step, we filled in a questionnaire with the information about our data and data organization in order to enable the Prague IT-experts to connect our data to "Manuscriptorium" later on.

In the middle of April 2008 the transformation of our data into METS format was successfully finished and so we agreed to do a first test. As a prototype I sent the XML-data in METS format (with embedded Dublin Core metadata) of the manuscript Cod. Pal. germ. 832 (http://digi.ub.uni-heidelberg.de/diglit/cpg832/), the so called "Heidelberg Book of Fate", to Prague.

Luckily there were no real technical problems regarding the processing of the Heidelberg METS-data for "Manuscriptorium". Everything was fine and clear: the available description, the detailed content overview and its relation to the pages and finally the pages list with URLs for all the available qualities.

To enable an easy way of harvesting and importing our data into "Manuscriptorium" the Heidelberg IT-department installed an OAI-interface within a short time so that Prague could and still can easily harvest the data as often as necessary. With the help of the OAI-interface my Heidelberg colleges were able to determine a so

called "set" so that Prague only harvested a special collection in our digital data, the digitized manuscripts of the Bibliotheca Palatina. Pretty soon a routine cooperation came into being and our data was inserted into the "Manuscriptorium/ENRICH" testing clone. But we still had to do some fine-tuning: e.g. the mapping of the data fields was improved and we wrote a summary of the conversion rules (end of May 2008). But at least everything happened within only a very few weeks and only by email contact.

Also our special wishes were regarded: By request Prague inserted a link to the original presentation of the digitized manuscripts on the Heidelberg server. For us this seemed to be the easiest way to show where the images are coming from.

At the end of May 2008 Prague was able to process 500 digitized manuscripts. In July 2008 the Heidelberg data was moved from the testing clone to the real routine "Manuscriptorium" database and in September 2008 the collaboration was announced on the official website of "Manuscriptorium/ENRICH".

Finally our directors signed the contracts. Hereby there was also arranged the access to the licensed documents of the other Manuscriptorium partners from Heidelberg campus for all users of Heidelberg University Library. Now they also could use the often licensed data of the other partners for free.

Meanwhile we're looking forward to continue our collaboration with "Manuscriptorium/ENRICH" also after we finished our project of the "Digitization of the German language Palatina Manuscripts".

# Session 4

# Future cooperation- beyond the European dimension

## Teuchos – A multilingual knowledge-based platform for research in classical philology

**Cristina Vertan**
Institute for Greek and Latin Philology, University of Hamburg, Germany

Abstract: Teuchos is a research infrastructure project aiming to to provide a web-based knowledge portal suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philological data. There are several challenges related with the implementation of the platform like:heterogeneity of the stored objects, on-line edition, multilinguality. In this article we will present a flexible architecture that tries to embed various types of objects a classical philologist would work with, link them and offer to the users cross-lingual services.
Keywords: *digital library, multilinguality, cross-lingual retrieval, ontology*

### Introduction

The development of web-based services opened new research facilities to paleography and codicology. Investigation on rare manuscripts, which until now were strictly dependent on the inspection of the physical object, can be done virtually, browsing digitized versions of the respective object. Moreover researchers can search through various data, compare assertions of different colleagues and come up with new theories regarding e.g. the origin the date or the provenance of a manuscript.

In comparison with digital libraries archiving modern documents, the objects in classical philology have particularities like:

- Are quite often only partially described,mainly due to the lack of information researchers have about manuscripts.

- It is almost impossible to define relations between objects which are valid for all elements inside a class
- One object contains text in several languages

Due to the above mentioned complexity up to now in classical philology we deal only with one object-type-repositories, which means that either it is a collection of manuscripts or a collection of watermarks, or collection of digitalized books, in the very best case with their descriptions (the most well known example is the Perseus digital Library [1])

The Teuchos Center for Manuscript and Text Research [2] was set-up in 2007 by the Institute for Greek and Latin Philology of the University of Hamburg in cooperation with the Aristoteles –Archive at the Free University Berlin. Teuchos is a long-term infrastructure project, which is financed in its starting phase (until mid-2010) by the German Research Foundation. In its final form Teuchos will to provide a web-based knowledge portal suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philological data.

In this article we will present a flexible architecture that tries to embed various types of objects a classical philologist would work with, link them and offer to the users cross-lingual services.

**Functionality**

The following use cases are foreseen for the research infrastructure:

- Provision of data facilitating the use of digitized manuscripts (created and shared by different user groups), ranging from structural information regarding the intellectual content of the manuscript to transcriptions containing indications of variant readings and eventually full-fledged digital editions.
- Provision of digitized manuscripts accompanied by (partial) transcriptions both as a basis for further editorial work and to make core information on the content and the manuscript tradition available to the scholarly community at the same early time.
- Collaboration of networked researchers independent of time and space as a prerequisite for the analysis and use of special materials;

An evolving collection of manuscript descriptions gives access to detailed information on codicology, manuscript history and textual transmission. This material derives from a library studies and is thus often inherently sporadic and disjointed; on

the other hand the collection is independent of library cataloging projects and open to the collaboration of researchers worldwide who contribute according to their respective field of expertise and/or their serendipitous findings.

A flexible model allows for the integration of manuscript descriptions of varying depth. A substantial amount of material taken from both published and unpublished materials of the *Aristoteles Graecus* [3] offers a model for comprehensive and highly structured descriptions.

All our objects are stored in a Fedora repository[1]. The user interacts with this repository via a web application that manages the editing, searching, and uploading processes. There are several groups of digital objects to be stored in the Fedora repository:

We store tracings of *watermarks* from dated paper manuscripts as digital images on the one hand, and descriptive data on these watermarks and their motif groups in an XML format on the other. Images are associated with Dublin Core[2] -like information about the data and linked to the descriptive metadata.

The *textual transmission* group is divided into two subgroups that are themselves subdivided: material related to individual *manuscripts* and material related to a particular *work*, e.g. a particular source text by a particular author.

The *manuscript* group encompasses digital *page images* of manuscripts (or parts of manuscripts) that are aggregated on a per manuscript basis scholarly *manuscript descriptions* that may reference page images if available for the one manuscript described, and *transcription* data, which may range from a first set of basic structural data to full transcriptions, and usually links to pages of exactly one manuscript (exceptions are e.g. texts spanning more than one manuscript volume and re- or misbound manuscripts).

The group of *works* encompasses a wide range of materials referring to a source text with its entire set of manuscripts rather than to one particular witness, and ranges from *full critical editions* (with several intermediate stages) and *translations* to various kinds of *commentaries* (and other explanatory or *descriptive* materials).

A special group is dedicated to *research papers* that may reference material from the other groups, without themselves falling into any of the other categories.

## Multilingual Aspects inside of Teuchos platform

Users of the Teuchos platform are speakers (or at least understand) one of five languages used inside of he community. At a first glance the straightforward consequence is the localization of the user interface in the envisaged languages. However, we claim in the

[1] cf. <http://www.fedora.info/>.
[2] cf. <http://dublincore.org/>.

following that there are deeper multilingual aspects which have to be handled and we illustrate how language technology can help.

We define three types of multilingual phenomena, occurring in our platform:

1) "***Macro-document***" - ***Multilinguality*** at the level of users and the uploaded multilingual documents:
   Therefore the platform requires not only to support for uploading of documents in all these languages but also to manage their relations to one or more manuscripts in a consistent way.
2) "***Micro-document***" - ***Multilinguality*** at the lever of primary data to be analysed. As we already mentioned manuscripts are accompanied by modern descriptions, critical texts, which although written in modern languages are containing often passages from the manuscript, or Latin citations. This is a real challenge when trying to process the documents automatically.
3) "***Terminological***" - ***Multilinguality***. related to watermarks. Even watermarks descriptions written in one language, may declare watermarks-motifs in a variety of languages. We have to ensure that watermarks are then classified as belonging to the correct class.

To handle these three types of multilinguality we propose an ontology based approach, integrating different ontologies related to components of the system. In each of the system main components (manuscripts, watermarks, etc.) a domain specific language independent ontology ensures the correct mapping of documents on the right concept(s). Links between components are realized between the nodes of the ontology and not the particulars instance objects (namely the documents).

## References


[1] Perseus, digital library, http://www.perseus.tufts.edu/hopper/

[2] Teuchos platform , http://www.teuchos.uni-hamburg.de/

[3] P. Moraux. *Aristoteles Graecus- d. griech. Ms. d. Aristoteles*. Berlin, New York: De Gruyter, 1976

[4] on-line Watermark collections, http://www.ksbm.oeaw.ac.at/wz/wzma.php, http://watermark.kb.nl, http://www.ksbm.oeaw.ac.at/wies/

# Digital Scriptorium: a Partnership

**Consuelo W. Dutschke**
Columbia University

**Abstract:** Digital Scriptorium is a growing partnership at present embracing twenty-eight American institutions. This paper touches upon six aspects of the partnership: its origin and its governance, its data collection and management, its image collection, and its funding. The URL for Digital Scriptorium is now http://www.scriptorium.columbia.edu but will change within the coming year as Digital Scriptorium moves its technology home back to its original base at the University of California, Berkeley.

## Digital Scriptorium: a Partnership

Politically, the United States began as a federation of independent colonies, and that historical approach to national unity manifests itself even in such small ways as a database of American-held medieval and Renaissance manuscripts. Digital Scriptorium is a voluntary consortium presently of twenty-eight partner institutions; it operates at the shared will of the group. It has no relationship with a centralized national government that imposes obligations on the nation's libraries to participate in DS; DS receives no regular funding from a governmental agency. Today I'll touch upon six aspects of the DS partnership: its origin and its governance, its data collection and management, its image collection, and its funding.

At its inception in 1997, DS was a joint program of the universities of Berkeley and Columbia; six years later it moved to Columbia alone; it is now returning to the University of California, Berkeley, but with the difference that executive power now formally lies not with the hosting university, but with a Board of Voting Members. These are the partner institutions that have accepted the responsibility of sending a representative to an annual meeting in which matters relevant to the group as a whole are decided. This shift in governance is organizational and will only function

successfully when there is openness in communication and cooperation between the host and DS. DS is not –yet, at any rate– a legally separate entity with its own tax status; it may eventually constitute itself in that manner; the advantage would be that it could, at that point, handle its own finances. The movement from casual cooperation, to written bylaws and annual meetings, to proposal for tax-exempt status as a legal entity are steps along a continuum, but with no radical change in the philosophy of the group.

The mechanism of data collection has passed through several stages. We had originally intended to use encoding in what at the time was SGML, now XML, but we were dependent upon another program completing its work on the development of a set of elements within an appropriate DTD. The work was eventually finished by a TEI Working Group under the designation, TEI-MS. Digital Scriptorium, in the meantime, because it was grant-funded with an upcoming deadline, and thus could not wait for the TEI product, determined to use as an interim solution a configured database in Microsoft Access. That Decision No. 1 had very long-range effects.

DS adopted Access as the standard data inputting tool, in part because it was widely known even by staff in very small libraries. In addition, we could easily impose controls in terms of required fields and data types, to ensure clean and matching data from the partner libraries. And Access, being a database, fit our goals: we did not aim for fully descriptive, finely nuanced presentations of the manuscripts: we intended the simplest, most naked description that was possible, that, together with images, would allow scholars to identify manuscripts crucial to their studies.

In addition, due to our grant funding and deadlines, we were obliged to make our results available immediately. Hence, Decision No. 2: we built our search application on top of the data architecture and field types of our Access database. Data aggregation, manipulation and indexing in fact take places in XML (we've used the open source platform, eXist, for this purpose), but reflect the datacentric origin of the material.

Until as recently as a few years ago, we still believed that our partners had the option of submitting their data in any number of formats. What changed our minds was that one partner submitted data encoded in TEI-MS, and another according to TEI P5. We paid for this diversity with so many hours of data massaging that this cannot be an option in the future. Instead, Berkeley, as the renewed host for DS proposes to implement inputting via a web-based interface to a MySQL database. Unified data, although originating in many different libraries, will be achieved via a simplified inputting tool, not via a multiplicity of supported formats.

DS recognizes that the more difficult it is for a library to participate, the less likely it is to do so. Therefore DS Central also accepts the duty of unifying diverse data via Browse Lists. One library may call a certain author "Peter Comestor," another may say

"Petrus Comestor," a third may input "Peter the Eater." To the search engine of a database, these are three different authors; DS Central unifies the various forms of a single author name (or title or scribe or artist) in its Browse Lists. When the user clicks on "Peter Comestor," entries originally input with any of the three forms will be retrieved.

The standards for images have always been fixed, although initially each partner hosted its own images. But not only could DS not count on 24/7 attention to server problems with the self-hosted images, but we were kept from implementing multi-resolution delivery, because we did not have the full body of TIFF images to submit to newer image-delivery software. We are now in the process of pulling all DS partner images together into a single repository. The ensemble of images will then also be retained, as a group, in dark storage for extra protection.

As you might expect from a program that is consortial in its other aspects, the financial situation of DS is also geared towards a consortial solution. The DS partner institutions are unequivocally committed to free service to our readers; we refuse user subscriptions because not only would users in non-subscribing communities be blocked from DS, but even the smaller contributing institutions would face a wall for viewing their own images. Therefore, if subscription is rejected, membership must take its place. DS is well positioned, better than most online academic resources, to look to its own members for financial sustenance, since there are large and growing numbers of DS partners: we have a committed body of stakeholders. The expense of running and developing DS, therefore, will be shared among multiple entities.

It must be said that DS is only at the beginning of sorting out the two complimentary financial issues: costs and income. The technology platform of the new host, Berkeley, is different from the one DS has lived with over the past six years, and we face certain one-time costs for migration of the data. On the other hand, the new technology will significantly simplify the aggregation and indexing of DS data. Berkeley is currently working on a budget to delineate the one-time versus the ongoing costs.

With regard to income, we are, I repeat, at the level of drafts and first discussions, so that I can only speak of plans, not fact. We have divided our participanting libraries into two main categories: academic libraries, and research institutions, and we've ranked each category into tiers. At present there are three tiers in each category with membership dues ranging from $2500 at the top, to $0 at the bottom; the DS participating members have ratified this first approach to sustainability.

We are confident that we are on a good path towards meeting our partner and user needs; our governance structure has proven itself for several years now; we are developing a system to address our finances. We are a changing work in progress.

Present URL: http://www.scriptorium.columbia.edu

**References available online 2009-09-15, listed in chronological order:**

Daniel V. Pitti, "Designing Sustainable Projects and Publications," in *A Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing, 2004 and online at http://digitalhumanities.org/companion/

*Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.* American Council of Learned Societies, 2006 and online at http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

Blue Ribbon Task Force on Sustainable Digital Preservation and Access,

*Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation: Interim Report* (2008) online at http://brtf.sdsc.edu/

Nancy L. Maron, K. Kirby Smith, Matthew Loy, *Sustaining Digital Resources: An On-the-Ground View of Projects Today*. JISC and Ithaka S+R, July 2009 and online at http://www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability/report/SCA_Ithaka_SustainingDigitalResources_Report.pdf

# The virtual manuscript room: looking beyond the single catalogue

**Peter Robinson**
University of Birmingham

It is a pleasant chance that the final conference of the ENRICH project should bring together myself and Consuelo Dutschke. We were the main architects behind the development of the two draft schemes for an XML standard for the encoding of manuscript descriptions which became the basis of the TEI-P5 formulation now used by the ENRICH project: myself, as leader of the EU MASTER project, and Consuelo, as co-leader, with Ambrogio Piazzoni, of the TEI workgroup.

However, neither of us, for various reasons, has had any involvement in ENRICH up to now. For myself, this gives me a rather unique perspective on ENRICH's extraordinary and ambitious attempt to build a Europe-wide catalogue. In the early stages of planning the MASTER project, we thought that, possibly, our work might lead to the creation of a single Europe-wide manuscript catalogue. We decided very early that this was an impossible and futile ambition. For many reasons, libraries and archives would want to keep various degrees of control over their records, as they do indeed over the manuscripts they keep. Accordingly, we decided to focus in MASTER on the encoding itself, with the idea in mind that if there could be a high degree of uniformity in the structure and semantics of the manuscript description records, then cross-searching of many records of many different repositories, held in many different systems, might be productive. Indeed, part of ENRICH has adopted this federated model, with great success: a glance over the partner list for ENRICH (as of late September 2009) shows some thirty six partner institutions have either joined or are in various stages of joining ENRICH.

This success might lead one to ask: were we wrong, many years ago, in concluding that there could not be a single Europe-wide manuscript catalogue?

ENRICH is now, by a very large distance, the single greatest resource we have in Europe for looking for manuscripts, and (by the magic of the digital world) displaying the pages. Parts of ENRICH go even further than this: including full text transcripts and translation of the manuscripts, as in the extraordinary Codex Gigas example. It is an great achievement, that ENRICH has got so far. Is it reasonable, to think that ENRICH could grow to become the single manuscript catalogue which we thought could never happen? Should it aspire to become this? Or even more, perhaps: as ENRICH can include transcripts, editions, translations and (we could expect) commentaries, analyses, the full apparatus of scholarship, might it not become the single home for the entire world of manuscript scholarship, and so far more than a catalogue?

I argue that the answer to these questions is, definitively, no. I go even further than this, and argue that the future for Europe-wide, and indeed worldwide, work on manuscripts does not lie with the single catalogue, hosted on a single site, which is part of ENRICH. Nor does it lie even with the federated catalogue, organized on an elaborate partnership of co-operating institutions, such as ENRICH has put in place (there is a similar, smaller, enterprise based on federated searching in the manuscript portal of the Consortium of European Research Libraries). These models all presume various kinds of institution-led initiative: consortia setting up and signing agreements, co-operations between teams of experts, agreements on protocols for exchange of data between partners. Such initiatives fit the funding model of various grant agencies very well (which is why we have seen so much money come into these initiatives), and they fit too the organizational model of the institutions very well too (which is why they have been so keen to seek funding). I have no doubt that this model will continue: the European Union and other agencies will continue to fund consortia projects with ever-increasing numbers of partners, offering to spread ever-wider nets of data across the web.

But for me, it is the wrong model. In the Virtual Manuscript Room at Birmingham, and with many partners across the world, we have been pursuing a different model. It is based on the perception that many manuscripts do not survive in well-funded institutions with the resources to join consortia such as ENRICH. It is based on the perception that most work on manuscripts –transcribing them, editing the texts from them, annotating them– is done by individual scholars working on their own. Consider the matter of manuscript images. In the ideal world, an institution would put the images on the web together with a full description, with transcripts, with introductions, with translations: such as has been done, for example, for the Codex Sinaiticus. We do not live in that ideal world. In our world (which is the only one we are likely to inhabit) a

library gets a bit of money to make a set of images of one of its manuscripts and puts them on the web, somewhere: for example, the 'Early manuscripts at Oxford University' project (http://image.ox.ac.uk/). Or, a team gets some money to arrange for a group of manuscripts to be photographed: like the 101st Airborne, they fly in and photograph all in sight in a frenzied few days. Some of the largest collections of manuscript images on the web have been made in this way, for example those at the Hill Museum and Monastic Library site (http://www.hmml.org/) and at the Centre for the Study of New Testament Manuscripts (http://www.csntm.org/). At these sites you will hundreds of thousands of manuscript images with virtually no information about the images or the manuscripts: typically, a note linking the set of images to a catalogue number, a sentence or two about the manuscript, and that is all. Around the world, there are many scholars who are interested in those manuscripts: who make lists of their contents, who transcribe the texts of the pages, and sometimes publish these, sometimes even as formal XML files, or as notes in blogs, or as emails, or simply leave these in their computers as Word or Excel documents.

Given this world, how do we best proceed? We think that the best route is to find ways to put all this information together. Traditionally, this is what catalogues did. But now, with the advent of web-wide tools, there is another route. This route is, simply put, the semantic web. This puts into the hands of every person with a computer the ability to find something on the web and then say something useful about it. Thus: a scholar in Sydney could look at an image on the Hill Museum website of a page from a manuscript in Armenia and say: that image of that page contains the first four verses of St John's Gospel. He or she could put that statement on the web in such a way that someone in Lyon could see it, in a few moments: and that Lyon scholar thinks: I will make a transcript of the text of that page. He or she then puts that on the web in such a form that in a few more moments, around the world hundreds of people interested in that text in that manuscript are reading the transcript.

An impossible dream? Many of us are dreaming this dream, across the world, now. And it is very possible, indeed. The proliferation of semantic web tools in the last decade, with their emphasis on precise, open and well-documented shared ontologies, provides one excellent starting point. The development of digital library systems over the same period, handling increasingly diverse data and offering increasingly powerful sets of tools, provides another starting point. Add to these the ferment of development of collaborating, lightweight services –seen first in 'social networking' sites and now being replicated across the academy– and the continuing immigration of scholarship into the digital world, and we have the ingredients we need. There will continue to be

catalogues, such as those of the great libraries, and such as that which ENRICH has made. But they will be part of a world of intelligent data, made by very many people and held on many different places, mostly outside the catalogues (even, outside any institution at all). As well as make the best catalogues we can: we can help make this other world of data as good as it can be.

# MULTILINGUALITY AND METADATA INTEROPERABILITY: THE CACAO PROJECT EXPERIENCE

**Luigi Siciliano**
University Library, Free University of Bozen-Bolzano, Universitätsplatz 1 - piazza Università, 1, 39100 Bozen-Bolzano, Italy

**Abstract:** The CACAO project is developing a system that will allow the user of libraries to type in queries in his/her own language and retrieve volumes and documents in any available language. In such a task two conflicting needs are strongly interrelated. On the one hand the need of comprehensive metadata in order to allow the Cross Language Information Retrieval System to work at its best, on the other hand the need of interoperability, in order to allow aggregation of catalogues of different institutions throughout Europe. This article describes the major issues and the solutions developed by the consortium in order to face these challenges, with particular regard to the implementation of two Application Profiles (AP) for Dublin Core Metadata. A Dublin Core Simple AP will allow the maximal interoperability and easy use whereas a Dublin Core Qualified AP – based on The European Library (TEL) Application Profile for Objects - will allow disclosing richer descriptive metadata and even enable each institution to define specific customizations according to local needs. From the technical point of view, the implementation of a modular XML schema has been a cornerstone towards the implementation of such infrastructure.
**Keywords:** *CACAO Project, cross-language, multilingual, interoperability, application profiles, metadata, Dublin Core, digital libraries, library catalogues.*

## 1. INTRODUCTION

The CACAO Project[1] addresses the need of access to multilingual resources in library catalogues by focusing on the query terms submitted by the users to the system that stores the descriptive metadata of these resources[2]. By coupling sound Natural

Language Processing technologies with available information retrieval systems, the CACAO Project aims at the delivery of a non-intrusive infrastructure to be integrated with current OPAC and digital libraries. CACAO therefore does neither translate descriptive metadata nor the resources itself, instead it translates and enriches the query. This will allow the user to type in queries in his/her own language and retrieve volumes and documents in any available language[3].

Several steps are required in order to obtain such a result.

First of all, the bibliographic records in the catalogues must be disclosed by each library and harvested by the CACAO system. The preferred way of harvesting metadata is via the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)[14]. In order to exploit sound Information Retrieval technologies (IR) and provide search results listed by relevance ranking, harvested data are off-line indexed .

As a second step, the query of the user is processed by three different subsystems in order to be analyzed (Part of Speech Tagging, or POS), translated and expanded (i.e. enriched with related terms) in all the available languages[4].

Eventually, the whole bunch of terms obtained at the end of the process is submitted to the search engine which will afterwards return search results sorted by relevance ranking.

Of course the task of automatically processing users' queries is quite complex because of many different matters, ranging from proper names identification to word sense disambiguation (WSD) and multiword issues.

The development of an Application Profile has been the most important tool in order to find the best solution for the first step, i.e. disclosing bibliographic records for harvesting.

[1] The CACAO Project (Cross-language Access to Catalogues and Online Libraries) is a 24-month targeted project supported by the eContentplus Programme of the European Commission[1, 4].
Members are Xerox Research Centre Europe as coordinator (FR), Free University of Bozen-Bolzano (IT) with both KRDB Research Centre of the Computer Science Faculty and the University Library, Gonetwork s.r.l. (IT), CELI s.r.l. (IT), Hungarian Academy of Science – Research Institute for Linguistics (HU), Cité des Sciences et de l'Industrie Library (FR), Goettingen State and University Library (DE), Kornik Library (PL), National Széchényi Library / MEK (HU).

[2] For the European context, according to a Eurobarometer Survey, "56% of EU citizens are able to hold a conversation in a language other than their mother tongue and 28% state that they master two languages along with their native language"[7, p. 8].

[3] Such a system may therefore address the needs of libraries and institutions operating in multilingual territories such as Switzerland or South Tyrol, or international federated catalogues, such as The European Library, which aggregates the catalogues of 46 of the 49 National Libraries of Europe[18].

[4] We are currently working on English, French, Polish, Hungarian, Italian and German. The whole architecture is based on web services, in order to achieve the highest modularity. Web services may be easily substituted or added according to the needs of different implementers.

## 2. Metadata and Multilinguality

The choice of OAI-PMH as preferred solution for harvesting data in CACAO is to some extent an answer to the issue of interoperability, since one key feature of this protocol is the requirement of Dublin Core Simple (DCS) as minimal common metadata format[6].

But DCS can be limited to the very basic descriptive information of the resource, and it can therefore be a useful resource for aggregating catalogues. However, it may lack useful indications for dealing with common problems of Natural Language Processing (NLP) and machine translation. For example, in the case of a query with the English term "Stove", a translation in German may be "Herd". "Herd" is a good translation but on the other hand the very same German word does have a different meaning in English. It is a typical case of a "false friend", i.e. words in different languages that are written in the same way but have completely different meanings. In such a case simply translating and forwarding the query to an index where terms are stored regardless to the language would lead to bad search results, retrieving not only cookbooks in German but also books in English dealing with sheep rearing[5].

Providing metadata with a clear indication of the language of the term via the xml:lang attribute[6] is a way of dealing with such an issue that can be pursued at the level of DCS[7]. This is why we developed a specific DCS Application Profile: even without addition of elements or attributes, recommendations and best practices for encoding DCS metadata may lead to a better quality of metadata.

## 3. Multilinguality and Interoperability

It is however true that richer metadata can lead to better results. The development of our Dublin Core Qualified (DCQ) format was supposed to address several needs. On the one hand, we had to satisfy some functional requirements for a better performance with the Cross Language Information Retrieval System (CLIR) in CACAO. On the other hand, four other important needs had to be taken into account.

[5] False friends and term ambiguity –with particular regard to the specification of the language of the metadata fields– have been examined in a recent paper presented by Barbara Levergood at the Dublin Core Conference 2008 in Berlin [12]. The example in the text is taken from this paper.

[6] For example:
<dc:subject xml:lang="de">Herd</dc:subject>
<dc:subject xml:lang="en">Stove</dc:subject>

[7] The identification of the language of a value can be achieved in two other ways: 1) infer the language from the values in other fields; 2) use language guesser software. These practices are however beyond the purpose of this paper.

The first was reusability, since CACAO libraries wanting to disclose metadata in a richer format also wanted to reuse it for other purposes in the future. Another important need was syntactic interoperability between partner libraries and The European Library (TEL)[18], as such integration would have been part of a specific Work Package. A third point was availability and easy implementation[8]. Last but not least, our AP needed to be flexible, in order to accommodate a range of libraries and records. The AP should allow a title and identifier as a minimum record, but also rich metadata supporting CACAO's NLP-based cross-language services.

The starting point has been the AP developed by TEL because it is an important precursor to Europeana[10], and exploiting its experience would have been a step in the right direction to a future-proof solution. TEL itself started developing its AP after 2001 using DC-Library AP as a basis[9]. Its metadata working group focused on both the collection level and the resource level, leading to two separated projects: TEL AP for Collection Descriptions[15] and TEL AP for Objects[16, 17]. TEL AP for Objects[10] closely stuck to DC Lib, adding as few elements as possible. CACAO AP maintained such structure, by giving a stronger status to specification of the language of terms and by focusing on Vocabulary Encoding Schemes (VES) as values for xsi:type attributes.

## 4. APs for both Human Beings and Machines: Recommendations and XML Schemas

Creating an AP which aims to be effective in the real world is a twofold problem: it should be understood both by human beings (e.g. librarians who prepare records for publishing) and software (e.g. client softwares that validate published metadata). Furthermore, it can neither be too vague nor impose detailed but unfeasible requirements for metadata.

With regard to the librarians, our AP clearly lists best practices and directions to be followed[11]. Each term is described in a distinct table, formatted according to the

[8] This ruled out several APs and encodings such as a Singapore Framework-based AP or The Europeana Semantic Elements (ESE), still under development, and Resource Description Framework (RDF) encodings, which currently many libraries would not be ready to implement.

[9] At that time the version used as basis was: <http://dublincore.org/documents/2002/04/16/library-application-profile/>. The current version is: <http://dublincore.org/documents/2004/09/10/library-application-profile/>. All Dc-Library APs share so far the status of Working Draft[5].

[10] The most relevant addition is the term <telRecordID>, added to allow identification of the item in the context of the collection[11, p. 40].

[11] Barbara Levergood played a major role in coordinating the activities in this Work Package. CACAO AP has been released as part of Deliverable 5.2 [13].

Dublin Core Application Profile Guidelines produced by the European Committee for Standardization (CEN) MMI-DC Workshop[8]. In addition to technical specifications, also more verbose explanations are provided, together with references to other related terms, thus making the implementation of each DC AP as painless as possible for system librarians.

In terms of software, XML Schemas[12] provide validation rules for well-formed documents by defining which elements and attributes are allowed in each namespace and which values can be used. To address the specific needs of a specific CACAO module for Word Sense Disambiguation (Word2Category)[2], a flexible and novel approach was used for the development of the XML Schemas. In fact, Word2Category relies on Classification System notations available in bibliographic records, requiring therefore in the dc:subject term a clear distinction between Classification Systems (CS) and Subject Headings (SH).

For this purpose a hierarchy of XML Schemas is used, allowing both a general distinction between CS and SH and the specification of a VES at a local level. The first is a general validation schema which includes the TEL schema and defines the general categories for CS and SH. Each library is free to use the aforementioned types or to define optional localizations (e.g. specify a CS such as *Regensburger Verbundklassifikation*) by means of a separated XML Schema file[13].

This kind of structure allows both interoperability, since all data are well formed, valid and easy to *dumb-down*[14], and adequacy, since distinction between CS and SH allows CACAO to work at its best and to exploit the Word2Category module.

## Conclusions

CACAO is a platform for cross lingual access to online catalogues and digital libraries. By developing Application Profiles for both Dublin Core Simple and

[12] XML Schema is the language for defining and validating an XML document supported by W3C: <http://www.w3.org/XML/Schema>. Other languages for the same purpose are Document Type Definition and RelaxNG: <http://www.relaxng.org/.>. However, the first is not an XML valid syntax and does have some limitations in validating the values provided in XML documents, whereas the second, although very powerful, is not a W3C Schema but is supported by OASIS, a Consortium of software vendors: <http://www.oasis-open.org/home/index.php>.

[13] From the technical point of view this can be achieved by using the attribute substitutionGroup. XML Schemas are available here: <http://www.unibz.it/library/standards/>. The XML Schema has been developed by Daniele Gobbetti of the KRDB Research Centre of the Faculty of Computer Science of the University of Bozen/Bolzano.

[14] For the dumb-down principle see: <http://dublincore.org/documents/usageguide/glossary.shtml#dumb>.

Dublin Core Qualified, CACAO partners have been able to provide libraries with directions for disclosing metadata that not only meet CACAO CLIR needs but are also reusable and interoperable with TEL. We are confident that the current interoperability with TEL will allow us to easier comply with forthcoming Europeana Semantic Elements (ESE)[15].

## References


Links and content checked: September 21, 2009.

1. Bernardi, R., Balestrieri, M., Bosca, A., Dini, L., Gobbetti, D., Segond, F. "CACAO System: An Overview". In *Proceedings of the Workshop on Advanced Technologies and Digital Libraries 2009. AT4DL 2009.* Bozen-Bolzano University Press : Bolzano, 2009, pp. 1-4, <http://purl.org/bzup/publications/9788860460301>.
2. Bernardi, R., Gobbetti, D., Siciliano, L. "Multilingual Access to Library Catalogues: Word Sense Disambiguation via Classification Systems". In *ICSD. International Conference for Digital Libraries and the Semantic Web. Proceedings*. University of Trento : Trento, 2009, pp.158-164.
3. Buoso, P., Siciliano, L., "Catalogo e ricerca multilingue: il progetto CACAO". In *Il mondo in biblioteca. La biblioteca nel mondo*, Editrice Bibliografica : Milano (In press).
4. CACAO Project, <http://www.cacaoproject.eu/ >.
5. Dublin Core Libraries Application Profile, <http://dublincore.org/documents/library-application-profile/>.
6. Dublin Core Metadata Element Set, Version 1.1, <http://dublincore.org/documents/dces/>.
7. European Commission, *Europeans and their Languages*, Special Eurobarometer 243 (2006), <http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf>.
8. European Committee for Standardization (CEN), *CWA14855 - Dublin Core Application Profile guidelines*, <http://www.cen.eu/cenorm/businessdomains/businessdomains/isss/cen+workshop+agreements/cwa14855.asp>.
9. Europeana Semantic Elements Specifications (v3.2), <http://version1.europeana.eu/web/guest/provide_content/>.


[15] Latest version is 3.2[9].

10. Europeana, <http://www.europeana.eu/>.
11. Levergood, B., Chambers, S., Siciliano, L. "Application Profiles Supporting Cross-Language and other Functionalities for Library Metadata". In *Proceedings of the Workshop on Advanced Technologies and Digital Libraries 2009. AT4DL 2009.* Bozen-Bolzano University Press : Bolzano, 2009, pp. 38-41, <http://purl.org/bzup/publications/9788860460301>.
12. Levergood, B., Farrenkopf, S., Frasnelli, E.: "The Specification of the Language of the Field and Interoperability: Cross-language Access to Catalogues and Online Libraries (CACAO)". In: Greenberg, J., Klas, W. (eds.) *Metadata for Semantic and Social Applications: Proceedings of the International Conference on Dublin Core and Metadata Applications 22-26 September 2008*, pp. 191-196, Universitätsverlag : Göttingen, 2008, <http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf>.
13. Levergood, B., Siciliano, L., Gobbetti, D., Dini, L., Bosca, A., Buoso, P., Barsanti, I.: *Integration with www.theeuropeanlibrary.org and aggregation of partner libraries. CACAO D5.2* (public), 2009, <http://www.cacaoproject.eu/outcomes/list-of-deliverables/>.
14. Open Archives Initiative Protocol for Metadata Harvesting, <http://www.openarchives.org/pmh/>.
15. The European Library Application Profile for Collection Descriptions (v1.5), <http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap_cld.html>.
16. The European Library Application Profile for objects (version 1.5), <http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap.html>.
17. The European Library Metadata Registry (for objects), <http://www.theeuropeanlibrary.org/handbook/regtable.php>.
18. The European Library, <http://www.theeuropeanlibrary.org/portal/organisation/about_us/aboutus_en.html>.

# APEnet Project:
## its impact on the european archives

**Luis R. Enseñat Calderón**

Abstract: APEnet stands for "Archives Portal of Europe in the Internet" and it is a consortium of twelve European state archives administrations, together with the EDL Foundation with two main objectives, the first one is the creation of an unique access point about the information contained in the European archives, and the second one to make this information consistent with Europeana and available thought it.

### Origin and legal framework

This is not the first attempt to create a unique access point to all the archival material in the Europe, but it is the more reliable and the one with more support of the European Union Institutions. The origin of the project dates back to 1991, but the 3 milestones of project are from 2003, 2005 and 2008.

In 2003, the Council of the European Union Resolution of 6 May 2003 on archives in the Member States invited the European Commission to submit to the Council of the European Union a report that would include orientations for increased future cooperation on archives at the European level.

Two years later, in 2005, at the request of the Council, a National Experts Group on Archives of the EU member States and EU institutions and organs elaborated a "Report on Archives in the enlarged European Union", that proposed five priority actions to increase archival cooperation in Europe, one of them was "The creation and maintenance of an Internet Gateway to documents and archives in Europe", this was the first time that the project was given an official name. The consequence of this report is the Council Recommendation of 14 November 2005 on priority actions to increase cooperation in the field of archives in Europe, published in Official Journal of the

European Union (29/11/2005) that recommends "the establishment and maintenance of an Internet portal for documents and archives in Europe" as a priority.

Once we have the legal framework for the project, the 3rd milestone in the implementation of the project. At the end of 2008, 12 European archival national administrations and the EDL Foundation presented a project to the European Commission e*Content*plus program in order to implement the Portal, signing in December 2008 a Grant Agreement to create and maintain the Portal.

### The final result

The project started in January 2009 and it is envisaged to create the first version of the portal at the beginning of 2011, and the final version the first days of 2011. As it is stated in the mentioned Grant Agreement, the overall goal of the APEnet project is to gather the existing digital archival content of Europe and make it available on-line, we do not plan to create new digital material, but to work with the existing one. The aim is to build a network of European archives that, can offer online access to finding aids covering digitised and not digitised documents, to the individual documents and digital objects through these finding aids, and information about individual collections, the institutions that house them, and their creators.

At the end of the project, information about 50.000 archival repositories, both private and public, will be available in the final portal, 16.000.000 multilevel descriptions of documents and archives and 31.000.000 digitised objects kept by these institutions. This huge amount of information will be available in Europeana too, but not all of it: in the Europeana portal the final user will be able to find the digital objects with its descriptions, mainly digitalised documents, but the information displayed in the APEnet gateway that will not have digital objects associated (documents that are not digitalised) will be only available throughout APEnet.

In the beginning of the project, the origin of the information will be the State Archives of Spain, Finland, France, Germany, Poland, The Netherlands, Latvia, Greece, Malta, Portugal, Slovenia and Sweden. But the archival materials are not exclusively in the custody of public archival institutions. In Europe, other institutions, like libraries and museums, house archival material, as is the case in the National Libraries of Spain and Malta or the British Library. Thus, the European Archives Gateway aims to facilitate the access to documents and records also in a variety of cultural heritage institutions, whether they are public or private.

Participation in the portal will be open to all archival repositories in Europe that can deliver structured descriptions of their holdings in accordance with international archival standards (either in EAD, EAC, EAG and METS format or in a format that can

be converted into EAD, EAC, EAG and METS with the help of converting engines provided by the project) In the project we do not intend to create new standards, but to follow the existing standards that are applied to Archives.

Many Member States have established national archives portals and gateways on the Internet based on these standards, sometimes with links to the individual records or documents. Built on the diverse archival traditions of the countries in Europe these portals and gateways are normally not conceived primarily to communicate and interchange data. The chosen standards are the Encoded Archival Description (or EAD) for encoding descriptions of finding aids, the Encoded Archival Context (or EAC) for encoding descriptions of record creators, the Encoding Archival Guide (or EAG) for encoding descriptions of archival repositories and the Metadata Encoded and Transmission Standard (or METS) developed to encode the structural metadata for digital objects and related descriptive and administrative metadata.

If one of the pillars of APEnet are the archival standards, the other one, as it is stated in the Grant Agreement with the European Commission, is the need to contextualise the content of archives holdings and collections in order to make individual archival objects searchable, accessible, and last but not least – usable. Most people can often on their own recall some aspects of the context of records and documents related to well-known persons or organisations. In order to reach a full understanding and use archival materials most effectively, however, they must be understood in relation to their provenance. The theoretical ground for this is the principle of provenance, which can be said to be the foundation of today's archival theory and practice, worldwide. In short, this principle states that an archival fond is the result of a records creator activity, developed step by step. The individual objects (records, documents) are parts of this process which can be fully reconstructed only with their help. The logical and physical place of each object mirrors its place in the process and defines its relations to other objects in the same process.

## Conclusions

The archival document is unique and seldom published. In most cases, the researcher must visit the archival institutions in person to access the material they contain. Public archival repositories in most member states and some private ones have already made multilevel archival descriptions and finding aids available on-line, to make it possible for the user to do research without knowing exactly where the sought-after information is physically located. The availability of on-line finding aids, especially if they are linked to the corresponding documents they describe, can save the researcher a

considerable amount of time and perhaps even eliminate the need to travel to the various institutions where the documents are housed.

To finalise, the European Archives portal can be described as a network of institutions that facilitates access to the existing archival resources across Europe, that contextualise the content of archives holdings and collections in order to make individual archival objects searchable, accessible, and last but not least – usable.

# Annex
# Content Partners
# Contributions

# DSP-Diocese Archives St. Pölten, Austria

The Diocese Archives St. Pölten contains a small stock of manuscripts and incunabula and furthermore coordinates and manages the most important digital library of European charters, Monasterium.Net.

Therefore the implementation to the Manuscriptorium platform in the frame of ENRICH comprises these three types of archival documents.

### Manuscripts and Incunabula

The collection of manuscripts covers 300 books from the beginning of the 13th to the 19th century –about 120 dating from the Middle Ages– and represents an important holding of the Diocese Archives St. Pölten. They originate from the former Augustinian monastery in St. Pölten, from several parishes and other monasteries in Lower Austria and consist especially of Biblica and Liturgica. Most of them are richly illustrated (e.g. Hs 1, Antiphonar from 1486, which contains illustrations of monks of St. Pölten) and therefore of historical and art historical value. Some of the books also contain Hebraic fragments (see the project of the Austrian Academy of Sciences: http://www.ksbm.oeaw.ac.at/hebraica/).

Furthermore the Diocese Archives keeps 386 incunabula and early printed books which mainly cover liturgical, historical, philosophical and canon law issues; they date from the 1470s to the 16th century. Over 270 incunabula are preserved in their original binding which make them singular for scientific studies.

The descriptions of the manuscripts and incunabula were available in Word-documents and therefore the bibliographical data were imported in Manuscriptorium via M-Tool. While digital images will only be allocated for the 120 medieval manuscripts, the metadata of all manuscripts and incunabula maintained in the archives will be integrated in ENRICH.

## Charters

The charter collections selected for ENRICH are part of a larger collection of charters within the Monasterium-Project (www.monasterium.net).

The approximately 45.000 charters with about 50.000 images to be part of Manuscriptorium originate to a great extent from monastery archives in Lower Austria, Vienna and Upper Austria but there are also charters kept by the state and federal state archives which come from former monasteries. They range chronologically from the 9th to the 18th century and are important sources for the early history of the mentioned regions. Detailed information on each object of the collection is available, including images and/or secondary data like text summaries, full texts and glossaries.

The integration to the Manuscriptorium portal was realised by developing an OAI interfaces for data harvesting and converting the CEI schema of the charters to the TEI-P5.

# BUTE-Budapest University of Technology and Economics National Technical Information Centre and Library, Hungary

**Dóra Emmert**

## From theology to legal studies, from profession to one's gratification – on the collection of BME OMIKK

Established in 2001 by the merge of two libraries with a history dating back to more than a 100 years, the National Technical Information Centre and Library of BME[1] is the largest source for natural and technical sciences in Hungary. The Library and Information Centre of the Budapest University of Technology and Economics was founded by the donation of Baron József Eötvös, the Hungarian Minister for Culture and Religion on the 9th of May, 1848. At first it only served the lecturers and students of the University, however, after the First World War it started to function as a public library as well. The National Technical Information Centre and Library has emerged from the library of the Museum of Technology and Industry, established by Ágoston Trefort, Minister for Culture and Religion on the 24th of June, 1883. The Museum undertook the task of maintaining a scientific library in order to provide sufficient material to readers with an industrial and technical interest.

Our library contributes to the Enrich project by supplying metadata of old and rare publications mainly in Latin, German and Hungarian. Amongst the collection are three incunabula –books printed before the 31st of December, 1500– in Latin and one in German. As for the old and rare books, six publications are presented from the 16th century, two from the 17th century, and the others were printed in the 18th century. Besides the books, we provide the metadata of articles from the first ever published architectural journal in German: Sammlung nützlicher Aufsätze und Nachrichten die Baukunst betreffend, Berlin, 1797-1806.

There is a wide range of subjects in the collection with which we have contributed to the Enrich project, including geometrics, arithmetics, astronomy, mining, agriculture as well as theology, chemistry, and mineralogy.

[1] Budapest University of Technology and Economics National Technical Information Centre and Library.

The following items may be of interest:

*From world history*

Hartmann Schedel: Register des Buchs der Croniken und Geschichten mit Figure und Pildnussen von Anbegin der Welt bis auf dise unsere Zeit[2] (Nürnberg, 1493) edition in German

Székely István: Chronica ez vilagnac yeles dolgairól[3] (Krakko, 1559) the first world history in Hungarian.

*On the subject of Caring For and breeding silkworm:*

Stephan Frendel: Die Kunst Seide zu erziehen…[4] (Bratislava, 1795)

Rövid oktatás az eperfák nevelésérŒl, és szaporításáról nem kölömben a selyem-eresztŒ bogaraknak hasznos tartásáról és az ugy nevezett galétának gyarapításáról[5] (Eszék, 1798)

*On hunting, falconry:*

Reliqua Librorum Friderici II. Imperatoris, de Arte Venandi cum avibuscum Manfredi Regis additionibus[6] (Augsburg, 1596);

Albertus Magnus: De falconibus, asturibus, et accipitribus[7] (Augsburg, 1596)

*On thermal waters in Hungary*

Torkos Justus János: Thermae Almasienses quoad earum situm…[8] (Pozsony, 1746);

Torkos Justus János: Schediasma de Thermis Pösthensiesibus[9] (Pozsony, 1745)

*Theology:*

Pázmány Péter: Hodoegus. Igazságra vezerloe kalauz[10] (Pozsony, 1637), a masterpiece of religious polemic writing from the era of counter-reformation.

*Cookery book from the 16th century:*

Marx Rumpolt: Ein neu Kochbuch[11] (Frankfurt, 1587)

[2] Chronicles and stories with illustrations, from the beginning of the world up to our time.

[3] Chronicle on notable events of the World.

[4] The art of breeding silkworms…

[5] Short instruction on the growing of mulberry trees and the culture of silkworms.

[6] The remains of the books of Emperor Frederick II. on the art of hunting with birds.

[7] On the falcons, the hawks and other birds of prey.

[8] Of the favorably located thermal waters of Dunalalmás.

[9] Thoughts on the thermal springs at Pöstyén (Hungary).

[10] A guide to Truth.

[11] A new cookery book.

*Medieval legal sources*

WerbŒczy István: Decretum oder Tripartitum opus, der LandtsRechten unnd Gewonheiten des Hochlöblichen Königreichs Hungern[12] (Wien, 1599) A German translation of the originally in Latin. The codification of Hungarian law served as his country's basic legal text until 1848.

We are glad to share metadata and images of the material mentioned above as we believe them to be beneficial for researchers, students or anyone who is interested in our cultural heritage.

[12] Decretum or Tripartitum opus, the right under the law of the Hungarian kingdom.

# ULW-University Library Wroclaw, Poland

**Grazyna Piotrowicz**

The University Library Wroclaw as the content partner in the ENRICH Project made much of a contribution to the development of Manuscriptorium Digital Library by providing access to many interesting manuscripts and old printed books from its special collections.

The digital collection of ULW provides access to reach collections of the manuscripts preserved in Department of Manuscripts (Special Collections Library located in the Library Building on the Sand Island in Wroclaw). There is the biggest collection of Silesian and Lusatian manuscripts in the world. Amount of the manuscript is: 12,532 library units, among them: Medieval manuscripts with fragments – about 3000, Oriental manuscripts– about 340, Greek manuscripts – 41, Cyrillic manuscripts– 11, Collection of the autographs – over 17,000.

The oldest fragment of the manuscript is from the 5th century (the fragment of the Chronicle of Eusebius, however the oldest codex is from the 9th century (Herbarium). The very important parts of that collection are illuminated codices, e.g. Psalterium nocturnum, Missale or Commentarius super Apocalypsim with the marvellous miniatures. Besides, there is a great collection of modern manuscripts (in Latin and German as well), among others can be found opus Topographia Silesiae by F. B. Wernher with many drawings of the monuments on Silesia from the 18th century. Our Library provided also in digital form the manuscripts from the former University Library in Frankfurt (this collection is called 'Viadrina') and other important items, like for ex. "Topographia Silesiae" by Wernher.

The digitized early printed books are from Department of Old Printed Books (Special Collections Library). This is the collection of more than 300,000 books, published from 15th till 18th century, including 3,200 incunabula, which is the largest collection of old books in Poland. They are of different provenance: the main part is the pre-war collection of the former City Library in Wroclaw and former University Library in Wroclaw. The large part of the books originates from different historical collections,

like: Bibliotheca Rudolphina, Bibliotheca Piastorum Bregensis and Bibliotheca Ecclesiana S. Petri et Pauli Legnicensis. The character of the collection is universal, it contains all the fields of science. They distinguish themselves by the large number of the prints, published by the most famous printers of Europe. Anyway the special feature of the collection is local, Silesian typographic production. So called Silesiaca are the rich and exceptional source for extensive scientific research on history of Silesia. Many old books present in Manuscriptorium are also Silesian books, e.g. Olsnographia by Johannes Sinapius, published in Frankfurt am Main in 1707 – the valuable source for studies on Silesian history, especially Olesnica. There is also another interesting book, written by the Silesian astronomer women, Maria Cunitz, published under the title: Urania propitia, published in Olesnica in 1650. The German-Polish manual from 1688 "Vierzig Dialogi" by Nikolaus Volckmar may be interesting for linguists. There is also a large number of occasional prints published because of funerals or weddings. They are also a valuable source for studies on history of Silesian families in 17th or 18th century. Within the framework of Enrich Project a large part of so called "Viadrina" collections are also digitised. These are old books from the former University Library from Frankfurt /Oder. Most of the digitized books are published in 17th or 18th century.

In Manuscriptorium there are also the manuscripts and old prints from Music Collection Department of ULW. The majority of them are from former St Elizabeth Church collection. There are the instrumental & vocal scores with religious contents written down by Johan Carl Poshner, who was a cantor in that church. It is a unique collection showing the reach music collection connected with the activities of church choir and instrumentalist bound at that church.

To Manuscriptorium Digital Library are provided also many old prints from Silesia-Lausitz Cabinet of ULW, where there are ca. 10,000 titles of so called Wratislaviana, i.e. collection concerning the history and life of city of Wroclaw. Majority of those old prints are unique and are the only source of history of the city. In the framework of ENRICH Project our Library ensure the access to 250 digital documents of that kind. There are songs of praise, occasional sermons, edicts of City Council, contents of homages paid to rulers and poetic works.

The way of technical cooperation of ULW with Manuscriptorium is described with details in Deliverable 5.3 of ENRICH Project.

# VUL-Vilnius University Library, Lithuania

**Elona Malaiskiene**

## Digital Collections of the Vilnius University Library

Vilnius University Library (VUL) is one of the oldest and richest academic libraries in Central and Eastern Europe. Established in 1570, it is a valuable resource both for the University and research community world-wide. Today VUL has a status of a research library of state significance with holdings of over 5,4 mln items. The most valuable collections (over 269 000 manuscripts and documents, over 172 000 rare books, 2237 old atlases and over 10 000 maps, collection of graphic arts of about 91 000 items etc.) are stored in specialized departments.

A small collection of parchments of the VUL Manuscript Department includes less than a hundred items. It include single land privileges or other privileges, land and estate selling documents, etc. signed by Lithuanian Grand Dukes and Polish kings. It also contains knighthood documents of individuals, popes' bulls and indulgences or assignments, documents of different monkhoods or their property, manuscripts of early European music

VUL Manuscript Department autograph collection consists of over 300 storage items. Collection embraces autographs of outstanding foreigners, famous rulers of Poland and Lithuania: Sigismundus the Old (1519–1526), Sigismundus Augustus (1562–1566), Stephan Bathory (1580–1583), and other kings, noblemen and their relatives; as well as Polish writers and public figures: A. Mickiewicz, J. Slowacki, and many others; French writers and artists, scientists: A. Decamps, Victor Hugo, P. Beranger, R. Chateaubriand, Voltaire, and others, Russian writers, historians, statesmen: G. Derzavin, F. Dostojevskij, A. Gercen, S. Glinka, I. Turgenev, Ekaterina II and others. Value of documents is not the same; however each document may be useful for researchers as an authentic material to witness the history or the fact of life.

Representatives of rich and influential noble families of the Grand Duchy of Lithuania, later Rzeczpospolita (from XVII c. also noblemen of Prussia) –the Radvilos

and the Sapiegos– served as the highest political statesmen –voivode, chancellor, hetman, marshal, Church officials. These noble families possessed castles, manors and residences all over the territory of the Grand Duchy of Lithuania. VU Library Manuscript Department holdings include archival material of their domains. It consists of documents of property managing, domain administration, economic activities legal processes, as well as personal and business correspondence; there are also documents related to their official responsibilities.

Collection of photographs consists of photos by Lithuanian photographers and photographers of other countries. Józef Czechowicz's photograph collection is especially important to Lithuanian culture history. He (about 1819–1888) is an outstanding Lithuanian photographer of the second half of XIX c. No less interesting is the Suprasl Orthodox Monastery photograph collection. Historians might be also interested in the Album of photographs from the Russian Tsar Nicolay II coronation ceremony in Moscow published in1896 by Polish photographer Jan Mieczkovski

V. Mincevicius (1915–1992) –a priest, journalist, translator and collector, who lived in Italy, donated his collection of maps to the University Library. It consists of 331 storage item. The greater part of his collection contains old cartography. They are items of great value of XVI–XIX c. world, Europe, regional maps and city plans created by world-famous authors and publishers of that time such as C. Ptolemaeus, W. J. Blaeu, G. Mercator, S. Münster, A. Ortelius, J. Hondius and others.

Herszek Leibowicz, a well known XVIII c.portrait engraver of Lithuania was an artist at Nieswiez, a Radziwill family castle in the Grand Duchy of Lithuania. Throughout 1745–1758 he forged in copper engravings, one hundred and sixty five portraits of the Radziwill family that were hanging in the art gallery of Nieswiez castle. VUL possess Radziwill portraits on separate pages published in Petersbourg.

All these digital collections will be presented in the Project.

ENRICH project is a wonderful chance to show the versatility of Lithuanian national cultural heritage held at Vilnius University Library to scientists and researchers all over the world –documents related to the history of Lithuania and Vilnius University.

# BNCF-Central National Library of Florence, Italy

**Pierantonio Metelli**

**The contribution to ENRICH of the Central National Library of Florence**

The Central National Library of Florence (hereafter referred to as BNCF) has its origins in the 30.000 volumes of the private library of Antonio Magliabechi, according to his will bequeathed in 1714 to the city of Florence. To increment the growing Library in 1737 it was decided by a mandatory decree that the new Library acquired a copy of all the publications printed in Florence and after 1743 in the entire Grand Duchy of Tuscany. In 1747 it was opened to the public for the first time with the name of Magliabechiana. In 1861 the Magliabechiana was unified with the Biblioteca Palatina (created by Ferdinand III of Lorraine and continued by his successor Leopold II) and assumed the name of National Library and from 1885 of Central National Library of Florence.

From 1870 any publication printed in Italy must be submitted to the BNCF by legal deposit. In its early days the Library had its headquarters in rooms belonging to the Uffizi and only in 1935 it moved to the present building. From 1886 to 1957 the BNCF published the "Bollettino delle pubblicazioni italiane ricevute per diritto di stampa", which in 1958 became "Bibliografia Nazionale Italiana" (BNI) (The Italian National Bibliography). The BNCF is also the pilot center for the creation of the National Library System (SBN), whose main aims are the automation of library services and the constitution of a national index of the collections of Italian libraries.

The main contribution provided by the BNCF to the ENRICH project was to supply digital contents of digitized manuscripts and books from its historical collections and to fix a technical framework to reach this goal (an ENRICH profile of metadata was created for the harvesting of data via OAI). The crucial core of this contribution can be identified in the Galileo Galilei's manuscripts, since it is a real unique collection in the history of science, world-wide studied and requested to the BNCF.

To summarize, the BNCF supplied the ENRICH project with the following digitized collections:

1) **Galileo Galilei manuscripts** (98650 images, 307 bibliographical units): almost the complete collection of Galileo Galilei manuscripts.
2) **Galileo Galilei printed books** (81678 images, 256 bibliographical units): books belonging to the private library of Galileo Galilei.
3) **Online manuscripts** (3865 images, 137 bibliographical units): the rarest and most consulted manuscripts owned by the BNCF (i.e. the Messale Ottoniano of the X Century, the Palatino 556, also named Lancelot, Filarete's treatises on architecture, etc.).
4) **Geographical maps** (3998 images, 947 bibliographical units): printed geographical maps, charts and military maps (XVII-XIX Century); handwritten maps and portolani (XV-XVII Century) and the handwritten maps of the cartographer Luigi Giachi (XVIII Century).
5) **Bibliotheca Universalis** (223361 images, 560 bibliographical units): manuscripts and printed books of English and French travellers in Tuscany (XVII-XIX Century), concerning topics related to the *Grand Tour* theme.
6) **Magliabechi** (211618 images, 52096 bibliographical units): partial digitizations (cover, title page, table of contents and variable significant pages) of printed books, mainly of XVI, XVII and XVIII century, for about 1/3 of the bibliographical units of the whole Magliabechi collection.

# BNE-Biblioteca Nacional de España

**Lourdes Alonso Viana**

### Contribution to ENRICH Project of the National Library of Spain

More than 300 digital objects are already accessible through ENRICH Project. We found among them some of the most important manuscripts of our library. It is expected this number will grow soon to more than 3000 with the addition of the complete collection of Incunabula and the autographs and first editions of the main plays of Spanish Golden Age theatre. This important achievement provides access to the documents held in one of the most valuable libraries in Western Europe.

The National Library of Spain was founded by Philip V in 1712. In 1836 it changed its denomination from Biblioteca Real to Biblioteca Nacional, and its management moved from the King to the Government. More than 70.000 documents came from the expropriation of the holdings kept in Convents, Churches and Cathedrals executed by Mendizábal in 1837. Some others came from different collectors. Nowadays, the main sources of the rise of the collection are auctions, direct acquisition from antiquarians, and donations.

The collection of the National Library of Spain is the country's most important one. Not only because it receives three copies of everything published in Spain by Legal Deposit, but, specially, because it holds the most of the national written heritage. One of the aims of the National Library is the dissemination of this heritage and the free access to the whole of the collection to researchers and users from everywhere in the world, anywhere they are.

Working with a parallel server to Biblioteca Digital Hispanica (institution's digital library), and with a conversion from our MARC21 records to the TEI scheme, some of the most valuable manuscripts are accessible through the ENRICH frame. It includes manuscripts often excluded from in-library use due to its value or to conservation reasons. We can find, for instance, the *Book of Hours* of Charles VIII, King of France; the will and testament of Elizabeth, the Catholic Queen; or the first epic poem in the Spanish Language, *Poema del Cid*, just to mention. Researches and scholars are able to

work on books that were only accessible through microfilm, slides or photocopies. The digital version provides a reproduction of more quality and free for the user who can study it from home in a more comfortable way. Besides, the scholar has not only the images but also a lot of tools for his/her study at his/her disposal thanks to the work made in user personalization.

Concerning the technical requirements, our way of cooperation was discussed directly with Tomas Psohlavec, from Aip Beroun. After completing a survey about the quantity and quality of the digitised documents available for the project, we worked on the way of sharing the data, and, as mentioned before, we created a parallel repository with the images accessible via ftp. The folder structure would have the images in one hand, and the XML descriptions in the other. This procedure allows the images to be open within the M-tool framework.

The documents available dates from the 1047 to the XIX Century, including an important collection of handwritten maps, but more documents are on the way, and the Incunabula will be soon included with text recognition added.

# NKP-National Library of the Czech Republic

The National Library of the Czech Republic has digitized its large historical collections since 1995 and therefore we can draw upon quite a large experience on this field. The Manuscriptorium system and its user interface was launched in 2003. The core of the Manuscriptorium digital library is created by the database of identification records. Nowadays there are about 180.000 of records fully accessible for public. The digital library consists also of six Thousands of fully digitized complex documents when some of them are supplemented by full text editions, and the amount is still growing. The access to the images is mostly licensed. The core of this virtual collection was created by the digitized collections of medieval manuscripts coming from the Czech National library. Among them we can find especially codices once possessed by the Charles University in Prague and by several important Bohemian monasteries. Nowadays there are also many digitized documents coming from other institutions from the whole Europe and even some Asian countries. Users can study all kinds of historical documents like illuminated manuscripts, Incunabula, Early printed books, historical maps, etc. The Manuscriptorium´s user interface provides various (simple or advanced) searching possibilities enabling relatively easy work with this historic material.

The natural challenge of such a kind of digital library is to integrate various types of sources from various institutions in single user interface. Therefore projects like the ENRICH project are so important for our work. It helps not only to enlarge the digital library and increase the number of provided digitized documents but it especially enables to establish a real cross-border cooperation and to start a real work on integration of various sources. Many concrete problems with aggregating of different formats (both of images and of metadata) could be (and has been) solved during the project.

The name of the project (ENRICH) was not chosen accidently. The final aim of such projects really is to *enrich* provided virtual research environment. A real enrichment is possible only thanks cooperation of institutions from various parts of

Europe what is the only way how to make accessible digital documents coming from various European regions. The easiness to study diverse sources normally being dispersed throughout the whole continent is one of the biggest advantages of Manuscriptorium digital library. Its virtual research environment opens new challenges for all scholars working with historical documents. Considerable speed-up of heuristic work and new searching possibilities enable to deal with themes and solve problems which would be unthinkable using classical research methods. The most marked examples of possible outcomes are various comparative studies overwhelming classic natural discourses and approaches. Besides these new study opportunities Manuscriptorium digital library very well presents the richness and variety of European written culture to wider public.

The ENRICH project belonged to rather larger projects on this field of activities. Works on the project joined together people from 18 partner institutions from 12 countries, set up links among them and created colorful international work group. Also the results of this effort are interesting and users of Manuscriptorium digital library will have the opportunity to gain from them very soon. We can just hope we will have the opportunity to continue with this common work to improve and enlarge it's fruits.

# KU-SAM-Nordisk Forskninginstitut at Copenhagen University, Copenhagen, Denmark and Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík, Iceland

## The Arnamagnæan Manuscript Collection

The Arnamagnæan Manuscript Collection derives its name from the Icelandic scholar and antiquarian Árni Magnússon (1663-1730) – Arnas Magnæus in Latinised form – who, in addition to his duties as secretary of the Royal Archives and, from 1702, professor of Danish Antiquities at the University of Copenhagen, spent much of his life building up what is by common consent the single most important collection of early Scandinavian manuscripts in existence, nearly 3000 items, the earliest dating from the 12th century. The majority of these are from Árni Magnússon's native Iceland, but the collection also contains many important Norwegian, Danish and Swedish manuscripts, along with about one hundred of continental European provenance. In addition to the manuscripts proper, the collection contains about 14000 Icelandic, Norwegian (including Faroese, Shetlandic and Orcadian) and Danish charters, both originals and first-hand copies (apographa).

The manuscripts are predominately written in Icelandic, Norwegian and Danish, with a smaller number in Swedish, Faroese, Latin, German, Low German, Dutch, Spanish, Italian and Basque. Vellum manuscripts make up about 20% of the collection, the remainder being on paper. Older bindings contemporary with the manuscripts themselves are few in the collection, and the majority of the manuscripts are in preservation bindings of recent date. Lavishly illuminated manuscripts are relatively rare in the collection (because rare among Scandinavian manuscripts generally), although some quite fine examples can be found among the manuscripts with religious or legal content.

Upon his death in 1730 Árni Magnússon bequeathed his collection to the University of Copenhagen where it was preserved until its division.

Even before its constitutional separation from Denmark in 1944 Iceland had begun petitioning for the return of the Icelandic manuscripts in Danish repositories and it was eventually agreed, in May 1965, that roughly half the items in the Arnamagnæan Collection (1666 items, in addition to all the Icelandic charters and apographa), should

be transferred to the newly established manuscript institute in Iceland, along with a smaller number of manuscripts (141) from the Royal Library (Det Kongelige Bibliotek) in Copenhagen. The first two manuscripts were handed over immediately after the ratification of the treaty in 1971 and the last two in June 1997, the entire process of transfer thus taking 26 years. The manuscripts transferred to Iceland have retained their original shelfmarks, and the two institutions which jointly act as custodians of the collection, the Arnamagnæan Institute (Den Arnamagnæanske Samling) in Copenhagen and the Árni Magnússon Institute for Icelandic Studies (Stofnun Árna Magnússonar í íslenskum fræðum) in Reykjavík, work closely together to ensure the long-term preservation of and access to the manuscripts in the collection.

### Nordisk Forskningsinstitut (Denmark)

Nordisk Forskningsinstitut (Department of Scandinavian Research), is part of the University of Copenhagen, Denmark. Its members of staff conduct research in the fields of Early Scandinavian language and literature, manuscript studies, Danish dialectology and socio-linguistics, onomastics and runology.

Den Arnamagnæanske Samling (The Arnamagnæan Institute) is a section within the Department. Its chief function is to preserve and further the study of the manuscripts in the Arnamagnæan collection. The academic staff of the section are responsible for research and instruction in the areas of Old Norse-Icelandic, Old Danish and Old Swedish, as well as Modern Icelandic and Faroese language and literature. Attached to the section there is a photographic studio and a conservation workshop, each with two full-time members of staff. The section publishes a series of scholarly monographs under the general title Bibliotheca Arnamagnæana and a series of critical editions of Old Norse/Icelandic texts, Editiones Arnamagnæanæ.

For the Enrich project 1601 images, taken from 16 manuscripts have been provided so far.

### Stofnun Árna Magnússonar í íslenskum fræðum (Iceland)

The Árni Magnússon Institute for Icelandic Studies is an academic research institute within the University of Iceland, operating on an independent budget and answering directly to the Ministry of Education. Its role is to:

- Conduct research on Icelandic Studies and related scholarly topics, especially in the field of Icelandic language and literature.
- Disseminate knowledge in these fields.
- Preserve and augment the collections within its care.

The 40.000 images from 400 of the manuscripts in the Stofnun Árna Magnússonar í íslenskum fræðum made available through the Enrich project represent only a part of the multi-faceted content of the Icelandic manuscripts in the collection. First and foremost the manuscripts of the famous Icelandic sagas are presented, with samples of manuscripts of other content.

# NULI-The National and University Library of Iceland

## The manuscript collection

The Manuscript collection holds the largest collection of Icelandic manuscripts to be kept in one place roughly 15,000 items.

The National and University Library of Iceland results from a merger from 1994 of the National Library, founded 1818 and the Library of the University of Iceland from 1940.

Now the leading and far the biggest library in Iceland.

The Library's manuscript department was originally a part of the National Library. A collection of at a least five generation of learned men, 3 bishops and a clergyman (the eldest born 1665) formed the basis for the National Library's manuscript collection. At the death of the bishop Steingrímur Jónsson in 1845, one of the founders of the National Library, this collection consisting of 400 manuscripts was bought from his family. Consisting of both their own production and their collecting of other manuscripts. In 1877 and 1901 the library purchased two more collcetion of 1337 and 1876 manuscripts respectively. The National Library's manuscript collection has grown steadily ever since to the 15000 items of today. Indvidual manuscripts and collections have been acquired through donation and, occasionally, purchase.

These manuscripts contain sagas, poetry, historical records, folktales, diaries and genealogical material both in original and copies. It represents the written part of the cultural history of Iceland. A lot of the material is what would be in printed books in other countries as the printing did not really have breakthrough in Iceland until in the middle of the nineteenth century. The tradition of copying became very strong in Iceland and the old literature the Sagas and alike of the golden age of Icelandic literature is preserved in numerous copies in the National libraries manuscript collection.

Paper manuscripts from the seventeenth, eighteenth and nineteenth century represent a very important element in the collection. The ones from the nineteenth century are most numerous. The twentieth century papers and letter collections and other personal papers from both societies and individuals, authors and other notable

people. Amongst important items from earlier centuries are five complete vellum manuscripts, some hundred vellum fragments and eighty two legal documents written on parchment. There is a small collection of pictures, photographs and paintings, although the accumulation of such materials does not represent a priority within the department.

One very special fragment should be mentioned. This is a leaf dated to around 1260 from a Norwegian Kings saga, Olaf the saint from the book Kringla by a thirteenth century scholar Snorri Sturlason. The manuscript was destroyed in a great fire of Copenhagen in 1728. This is the only leaf to survive and was in a mysterious way kept in the Royal Library in Stockholm and presented to the Icelandic people in 1975 by the Swedish king.

The images of manuscripts The National and University Library in Iceland adds to the Manuscriptorum portal from its collection to be available through the Enrich project are all from the category of the Icelandic sagas. The heroic and legendary family sagas of Icelanders, the settlers and the first few generation written mainly from the end of the 12th century to the end of the 14th century, being preserved in endless copies down to the twentieth century along with the printed versions.

# CSH-Cologne MNS

## The Early Printed Books Provided to ENRICH by the Forschungsarchiv für Antike Plastik, Cologne

The corpus of early printed books on the subject of archaeology and the classics provided by the Forschungsarchiv für Antike Plastik and its database project Arachne for the ENRICH project stems from three different sources and is by no means complete, as it will grow in the next two years to thrice the size which is now available via ENRICH.

Begun in 2006, the Forschungsarchiv has a cooperation with the Winckelmann Society and the German Archaeological Institute in Rome to digitize and make available all the early printed books in these collections. Originating in the close partnership between the Forschungsarchiv and the Chair for Computer Science for the Humanities, which is an ENRICH partner, the Forschungsarchiv decided to make this ongoing effort available for harvesting by ENRICH.

The books are all fine examples on the reception of antiquity in the 16th to 17th century, and most of them are artfully illustrated. All in all, the Forschungsarchiv now (Oct. 2009) provides 300 early printed books with around 46'000 pages with an additional 600 books with around 100'000 pages to come in the next two years.

## The Rara Library of the Archaeological Institute at the University of Cologne

As a young archaeological institute at a german university, the library and the collection of rara present in Cologne had to be brought retrospectively after the creation of the institute in 1928.

Where old institutes like Göttingen or Bonn were able to buy the books at the time they actually were imprinted, the Cologne institute had to get theirs from auctions, book dealers and collection sales. This had one advantage: most of the books are on the focal matters of the Cologne institute: collection history, roman scuplture and architecture, topography and reception history.

### The Collection of the Winckelmann Society

The Winckelmann Society's main focus of research is the life and works of Johann Joachim Winckelmann (1717-1768), the founder of scientific archaeology. Most of the rare books in the collection of the Society deal with art and architecture which Winckelmann himself had known and the imprints are often contemporary or shortly after the life of the patron of this collection.

### The Collection of the German Archaeological Institute in Rome

The library of the German Archaeological Institute in Rome is the biggest archaeological library in the world. It houses more than 210'000 volumes on the subject of archaeology and related disciplines, as well as incorporating the Biliotheca Platneriana, whose main interest lies in the development of italian towns. The library has a very rich collection of books on archaeology from the 16th to the early 19th century, which is being subsequently digitized.

ENRICH PROJECT

EUROPEAN NETWORKING RESOURCES AND INFORMATION CONCERNING CULTURAL HERITAGE