DigItalia

Anno V, Numero 2 - **2010**

Rivista del digitale nei beni culturali

ICCU-ROMA

The ENRICH project: Towards a European digital manuscript library

Matthew James Driscoll
The Arnamagnæan Institute; University of Copenhagen

Background

The idea of using computers to provide greater access to medieval manuscripts and other primary sources dates from the late 70s and early 80s, when a number of attempts were made to apply relational database technology to manuscript studies, in particular in the form of searchable electronic catalogues. Unfortunately – but understandably – these projects generally relied on locally developed or proprietary software, with all the problems for long-term maintenance and interoperability that entails. Moreover, each system tended also to have its own standards with regard to the nature, extent and organisation of information included, reflecting the lack of often even national standards for manuscript description at the time.

In the mid-Nineties the advent of Standard Generalized Markup Language (SGML) and the World Wide Web gave new impetus to work on electronic manuscript cataloguing. At the same time, developments in digital imaging meant that manuscript holding institutions could provide an unprecedented degree of access to their holdings. With the rise of large-scale digital collections came an increased awareness of the central importance of metadata standards.

In November 1996 a meeting was held at Studley Priory, near Oxford, organised by Peter Robinson of de Montfort University and Hope Mayo from the Mellon-funded EAMMS project (Electronic Access to Medieval Manuscripts) and attended by representatives from major manuscript holding institutions in Europe and the United States, together with experts on MARC, the Berkeley Finding Aids project, the TEI (Text Encoding Initiative) and Dublin Core. A year later there was a similar meeting at Columbia University in New York which brought together many of the participants in EAMMS, Digital Scriptorium (also funded by the Mellon Foundation) and several other manuscript-related projects. These meetings, both attended by the present writer, confirmed that there was indeed not only a widespread awareness of the need for an international standard for manuscript description, but also a fairly broad consensus as to what form that standard should take and what the appropriate technical means were to implement it, viz. something alsong the lines of the *Guidelines for Electronic Text Encoding and Interchange* developed by the TEI, an international and interdisciplinary standards project established in 1987 to develop, maintain and promulgate hardware – and

software – independent methods for encoding humanities data in electronic form[1]. In 1999 funding was obtained from the Telematics for Libraries section of the European Union Fourth Framework research programme for the establishment of the MASTER project (Manuscript Access through Standards for Electronic Records), whose goal was to define and implement a general purpose standard for the description of manuscript materials using TEI-conformant XML[2]. The project ran through 2001 and was, by the standards of many EU-funded projects, reasonably successful, in that the system it developed was actually adopted by many large-scale electronic cataloguing projects. Among the largest and most important of these is *Manuscriptorium*, a digital library of manuscripts and early printed books developed and maintained by the Czech National Library in Prague[3].

The most recent version of the TEI *Guidelines*, TEI P5[4], released in November 2007, contains a major new chapter on manuscript description which is based largely on the work of the MASTER project and the TEI Medieval Manuscripts Description Work Group (TEI-MMSS), active between July 1998 and October 2000, which was headed by Consuelo Dutschke of the Rare Book and Manuscript Library, Columbia University, and Ambrogio Piazzoni of the Biblioteca Apostolica Vaticana. Although the work of these two groups proceeded largely in tandem, and despite an avowed intention that a single set of recommendations should emerge from them, there were, in the end, some significant discrepancies between the two proposed schemes. The MASTER project, for example, never finalised its discussion on seals before the end of the project period, while TEI-MMSS did, whereas MASTER developed quite sophisticated mechanisms for dealing with bibliographical and prosopographical data, an area largely untouched by the Work Group. In this sense the two schemes could be said to complement each other. There were, however, also discrepancies between the two which seemed to reflect a fundamental difference of opinion as to what the system should be used for and by whom. Thus TEI-MMSS, which consisted principally of librarians and cataloguers, seemed primarily concerned with the practicalities of manuscript cataloguing, and in particular with the accommodation of existing (legacy) data, while the MASTER project, which consisted principally of manuscript scholars and mark-up experts, seemed more interested in determining the underlying structure of manuscript descriptions in a more general, theoretical way. In order to resolve this is-

[1] For information on the TEI see http://www.tei-c.org.

[2] Principal project members were The Centre for Technology and the Arts at De Montfort University, Leicester (UK), Oxford University's Humanities Computing Unit (UK), Koninklijke Bibliotheek, Den Haag (NL), L'Institut de recherche et d'histoire des textes, Paris (FR), Národní knihovna âeské republiky, Praha (CZ) and Det Arnamagnæanske Institut, København (DK). Unfortunately, the MASTER website was not maintained after the end of the project, but a number of cached copies of MASTER-related documents can be found on http://xml.coverpages.org/master.html.

[3] http://www.manuscriptorium.com.

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

sue, the TEI Council in 2002 appointed a special task force, led by the present writer, whose job it was to review the current state of TEI-based recommendations for the detailed description of manuscript materials and define a common subset of those recommendations adequate to the needs of the TEI community. Because the task force was able also to take into account the actual experience of the many electronic cataloguing projects then under way, the manuscript description module eventually incorporated into TEI P5 is not simply a common subset of the two schemes, but rather a significant improvement on both[5].

ENRICH

In 2007 the ENRICH project received funding under the eContent*plus* programme with the aim of extending *Manuscriptorium* to create seamless access to distributed information on manuscripts and early printed books throughout Europe, while at the same time upgrading the underlying metadata from Master*plus* (essentially the MASTER standard with added structural metadata) to TEI P5[6]. One of the project's central work packages, WP3, dealt with the "standardisation of shared metadata". Its goal was to ensure interoperability of the metadata used to describe all the shared resources by analysing the various standards used by different partners and ensuring their mapping to a single common format, which will be expressed in a way conformant with current standards.

The first thing that was done within the project was therefore to assess what differences there actually were between TEI P5 and Master*plus* and then resolve these differences. A wide sample (more than 1.000) of existing manuscript description records in many formats was reviewed, allowing the identification of a common core of practice. On the basis of this, a narrow subset of the TEI – which is designed to support a huge range of document types and encoding practices – was defined, including 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.

Other constraints were added, for example, by pre-defining the contents of many attribute value lists rather than leaving them open, and making a number of attributes obligatory rather than optional.

[5] For more information on this process see my article *P5-MS: A general purpose tagset for manuscript description*, http://www.digitalmedievalist.org/journal/2.1/driscoll/.

[6] Partners in the ENRICH project were: Národní knihovna âeské republiky, Praha (CZ), AIP Beroun, s.r.o., Beroun (CZ), Oxford University Computing Services (UK), Centro per la comunicazione e l'integrazione dei media, Università degli Studi di Firenze (IT), Matematikos ir informatikos institutas, Vilnius (LT), SYSTRAN s.a., Paris (FR), Biblioteca Nacional de España, Madrid (ES), Biblioteca Nazionale Centrale di Firenze (IT), Vilniaus universiteto biblioteka (LT), Biblioteka Uniwersytecka we WrocÏawiu (PL), Stofnun Árna Magnússonar í íslenskum fræ?um, Reykjavík (IS), Universität zu Köln (DE), Monasterium Projekt, Diözese St. Pölten (AT), Landsbókasafn Íslands – Háskólabókasafn, Reykjavík (IS), Budapesti Mıszaki és Gazdaságtudományi Egyetem (HU), Poznaƒskie Centrum Superkomputerowo-Sieciowe (PL) and Den Arnamagnæanske Samling, Nordisk Forskningsinstitut, Københavns Universitet (DK).

Let us look at one example. The TEI manuscript description module defines a number of specific elements designed to contain certain types of information. These are:

- <msIdentifier>: groups information uniquely identifying the manuscript, such as holding institution and shelfmark;
- <msContents>: provides an itemised list of the intellectual content of the manuscript, with transcriptions of rubrics, incipita, explicita etc., as well as primary bibliographic references;
- <physDesc>: groups information concerning all physical aspects of the manuscript, its material, size, format, script, decoration, binding, marginalia etc.;
- <history>: provides information on the history of the manuscript, its origin, provenance and acquisition by its current holding institution;
- <additional>: groups other information about the manuscript, in particular administrative information relating to its availability, custodial history, surrogates etc.

Within each of these, further specialised elements are defined. The <physDesc> element, for example, can contain elements for describing features such as the nature of the support, the dimensions of binding, leaves and written area, the foliation, pagination and columnation, the collation or quire structure, the layout of the page, the scripts used and identification of the hands, of known, as well as descriptions of illumination, decoration, paratextual features, musical notation etc. Use of all of these elements, apart from <msIdentifier>, is optional in the TEI, and often there is more than one possible way to provide the same information. For the purposes of the ENRICH project, however, it was decided to make a large number of elements and attributes obligatory, in order to ensure that all partners provided at least some basic types of information and encoded it in the same way. In order to indicate the nature of the support, for example, it was decided that the @material attribute on the <supportDesc> element should be compulsory, and that it must take one of the following values: "perg", for parchment, "chart", for paper, "mixed" or "unknown". In this way, the support is given for every manuscript in the system, and in a way which is searchable regardless of the language in which the manuscript description is written[7].
Synchronising ENRICH's requirements with TEI P5 necessitated close collaboration with the TEI Council, which was revising the manuscript module at the same time. It was also important to work closely with "AiP Beroun", the private firm who acted has technical co-ordinator for the project, to ensure that the Manuscriptorium platform would in fact be able to support the full complexity of TEI P5. Finally, it was necessary that a complete consensus among partners was reached.

[7] For more information on the relationship between the ENRICH project and TEI P5 see http://enrich.manuscriptorium.com/index.php?q=node/9.

The ENRICH standard was formally defined using TEI ODD (One Document Does it all) – the source format in which the *TEI Guidelines*, including the schema fragments and prose documentation, are written in a single XML document – which allows the automatic generation of schemata in DTD (Document Type Definition) and the RelaxNG (Regular Language for XML Next Generation) and W3C (World Wide Web Consortium) XML schema languages, as well as full documentation in a variety of languages (French, Italian, Spanish and English). The ENRICH standard has been tested in many different training contexts and a suite of training materials produced, covering the basic ideas of XML markup as well as the TEI modules for metadata, basic document structure, manuscript description and transcription, persons and places, facsimiles and non-standard writing systems[8]. A suite of XSLT (eXtensible Stylesheet Language Transformations) stylesheets and associated workflows – collectively known as the "ENRICH Garage Engine" – has also been developed for conversion from existing metadata formats such as EAD (Encoded Archival Description), MASTER and MARC (Machine Readable Cataloging), while the ENRICH "Gaiji Bank" is a tool for dealing with non-standard characters and glyphs, something which is often crucial for those working with manuscripts and other historical documents[9].

In sum, ENRICH provides a system which facilitates both the lossless conversion of existing manuscript description data and the creation of completely new data. What is more, ENRICH can be used to produce the complete digital surrogate, comprising a collection of digital images of the manuscript, an associated TEI Header – the metadata component of any TEI document – containing a description of the manuscript, an encoded transcription of the manuscript's text(s), optionally incorporating layers of scholarly interpretation and analysis, and an associated body of factual information about e.g. the persons, places, organisations and events related to the manuscript – and link all these components seamlessly together.

Handrit.org

The way in which this works in practice can be seen from handrit.org., a digital library of Icelandic manuscripts, which is a collaborative effort by three partners in the ENRICH project, the Arnamagnæan Institute (Den Arnamagnæanske Samling) in Copenhagen, the Árni Magnússon Institute for Icelandic Studies (Stofnun Árna Magnússonar í íslenskum fræðum) in Reykjavík and the National and University Library of Iceland (Landsbókasafn Íslands – Háskólabókasafn).

Handrit.org was conceived as a central point of access for information about and analysis of the manuscripts in these three collections, which between them com-

[8] Links to all these documents and tools can be found on the OUCS website: http://tei.oucs.ox.ac.uk/ENRICH/.

[9] For the ENRICH Garage Engine see http://dl.psnc.pl/software/EGE/; for the Gaiji Bank, see http://manuscriptorium.com/index.php?q=gaijibank.

prise nearly 90% of the Icelandic manuscripts extant[10]. 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 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. 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. 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.

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