# THE EXPERIENCE OF THE ARNAMAGNÆAN INSTITUTE, COPENHAGEN

Matthew Driscoll

*Abstract: The Arnamagnæan Institute, principally in the form of the present writer, has been involved in a number of projects to do with the digitisation, electronic description and text-encoding of medieval manuscripts. Several of these projects were dealt with in a previous article 'The view from the North: Some Scandinavian digitisation projects', NCD review, 4 (2004), pp. 22-30. This paper looks in some depth at two others, MASTER and CHLT.*

*The Arnamagnæan Institute is a teaching and research institute within the Faculty of Humanities at the University of Copenhagen. It is named after the Icelandic scholar and antiquarian Árni Magnússon (1663-1730), secretary of the Royal Danish Archives and Professor of Danish Antiquities at the University of Copenhagen, who in the course of his lifetime built up what is arguably the single most important collection of early Scandinavian manuscripts in the world, some 2,500 manuscript items, the earliest dating from the 12th century. The majority of these are from Iceland, but the collection also contains important Norwegian, Danish and Swedish manuscripts, along with approximately 100 manuscripts of continental provenance. In addition to the manuscripts proper, there are collections of original charters and apographa: 776 Norwegian (including Faroese, Shetlandic and Orcadian) charters and 2895 copies, 1571 Danish charters and 1372 copies, and 1345 Icelandic charters and 5942 copies. When he died in 1730, Árni Magnússon bequeathed his collection to the University of Copenhagen. The original collection has subsequently been augmented through individual purchases and gifts and the acquisition of a number of smaller collections, bringing the total to nearly 3000 manuscript items, which, with the charters and apographa, comprise over half a million pages.*

## Projects

Following its constitutional separation from Denmark in 1944, Iceland petitioned for the return of Icelandic manuscripts in Danish repositories. After much debate, it was finally agreed that a significant part of the Arnamagnæan Collection (1666 items, in addition to all the Icelandic charters and apographa), should be transferred to Iceland, along with a smaller number of Icelandic manuscripts (141) from the Royal Library in Copenhagen, to be housed in a sister institute set up expressly for that purpose. The first two manuscripts were handed over in 1971, immediately after ratification of the treaty, and the last two in June 1998. At about that same time the Arnamagnæan Institute (in close cooperation with its sister institute in Iceland) began working towards reuniting the collection virtually through digital technology. Outlined below are some of the projects and initiatives in which the institute has become involved as a result.

*The MASTER project and the TEI*

The first step toward the goal of virtual reunification will be a new electronic catalogue of the entire collection, based on Kristian Kålund's Katalog over Den Arnamagnæanske Håndskriftsamling (Copenhagen, 1888-1894), but supplemented by more recent scholarship. Because the two Arnamagnæan Institutes are primarily research institutes, whose chief function is to further the study of the manuscripts in the collection, our records tend to be fuller than most ordinary catalogue entries, and contain occasionally quite detailed descriptions of palaeography and orthography, illumination and bindings, as well as full transcriptions of marginalia and accompanying material, such as the notes made by Árni Magnússon and his amanuenses, which are generally kept with the manuscripts to which they refer.

Preliminary work on this catalogue was undertaken as part of MASTER (Manuscript Access through Standards for Electronic Records), an international project with funding from the Telematics for Libraries section of the European Union Fourth Framework programme whose goal was to define and implement a general purpose standard for the description of manuscript materials using TEI-conformant SGML/XML (the project website, unfortunately now rather out of date, is: http://www.cta.dmu.ac.uk/projects/master/). The project period ran from January 1999 through June 2001. Full partners, in addition to the Arnamagnæan Institute, were: Centre for

Technology and the Arts at De Montfort University, Leicester (UK), Koninklijke Bibliotheek, Den Haag (NL), L'Institut de recherche et d'histoire des textes, Paris/Orleans (FR), The Humanities Computing Unit, Oxford (UK) and Národní knihovna České republiky, Praha (CZ). Associate partners included several major European libraries, notably The British Library (UK) and Biblioteca Apostolica Vaticana (VA), as well as a number of smaller institutions such as our sister institute in Iceland, Universitetsbiblioteket, Lund (SE), Народна Библиотека "Св Св Кирил и Методий" and Институт по Математика и Информатика, БАН, София (BG), and Lietuvos nacionaline Martyno Mazvydo biblioteka, Vilnius (LT). An independent expert group, made up of Dr Ian Doyle, Durham (UK), Professor Peter Gumbert, Leiden (NL) and Dr Gilbert Ouy, Paris (FR), monitored and commented on the development of the standard from the start. Since the end of the project period there has also been significant input from users of MASTER, which number in the hundreds, if not thousands.

MASTER had close contacts with several other projects with similar or complimentary goals: in North America the EAMMS project (Electronic Access to Medieval Manuscripts), a collaboration between the Hill Monastic Manuscript Library at Saint John's University in Minnesota and the Vatican Film Library at Saint Louis University, funded by the Andrew W. Mellon Foundation, and Digital Scriptorium, a collaboration between the Bancroft Library at the University of California at Berkeley and Columbia University's Rare Book and Manuscript Library, also funded by the Mellon Foundation; and in Europe MALVINE, funded under the same EU call as MASTER, but focusing on modern literary manuscripts and letters. Finally, the development of the MASTER document type definition (DTD) for manuscript description proceeded in tandem with the Text Encoding Initiative's Medieval Manuscripts Description Work Group (1998-2000), chaired by Consuelo Dutschke, Curator of Medieval and Renaissance Manuscripts at the Rare Book and Manuscript Library at Columbia University and Ambrogio Piazzoni, prefect of the Vatican Library.

More or less as a direct result of this, the institute became a member of the Text Encoding Initiative itself. The TEI is 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. The TEI began as a research effort cooperatively organised by three scholarly societies (the Association for Computers and the Humanities, the Association for Computational Linguistics and the Association for Literary and Linguistic Computing), and funded by substantial research grants from, among others, the US National Endowment for the Humanities, the European Union, the Canadian Social Science Research Council and the Mellon Foundation. In December 2000 an independent and self-sustained non-profit consortium was set up to maintain and develop the TEI standard. There are currently 81 members of the TEI Consortium, including universities, research libraries, academic and other non-profit publishers, scholarly societies and others concerned with the design, production or delivery of structured electronic text (see the TEI's website: http://www.tei-c.org). The technical work of the TEI is overseen by an elected council, on which I have served sinced 2000.

The next version of the TEI Guidelines, "P5", scheduled for release in the early part of 2005, will contain a major new chapter on manuscript description, based chiefly MASTER and the work of the TEI workgroup, but with significant input also from the Repertorium of Old Bulgarian Literature and Letters project, based in Sofia and Pittsburg. I chaired the TEI task-force whose job it was to reconcile these various schemes and produce a single system for incorporation into the TEI. There remains a number of areas in need of further work, however, which will be dealt with by a properly constituted work-group, of which I shall probably also be the chair.

Meanwhile, work on the Arnamagnæan catalogue continues. During the MASTER project period itself some 500 records, the majority of them minimal, were produced in Copenhagen. It was decided to concentrate on the medieval manuscripts in the collection, although post-medieval manuscripts of special importance (for example copies of medieval vellums now lost) were also described. Since the end of the period minimal records – comprising little more than shelfmark, date and place of origin and an identification of the contents – were made for the remainder of the collection, but little more than that has been done owing to lack of funds. In Iceland basic cataloguing began in the year 2000. It was decided to include all information regarding each manuscript from the printed catalogue, translated from the original Danish to Icelandic, but, in the initial stages, no more than that. Two full-time employees carried out most of this work. All the manuscripts in the Icelandic part of the collection have now been catalogued in this manner and work has begun on "complete cataloguing", where each manuscript is examined and its contents and appearance described in detail.

The cataloguers at both institutes have produced manuals outlining the methods and terminology for cataloguing. Furthermore, the cataloguers in Iceland have in cooperation with the Icelandic software company Raqoon

designed a markup language for manuscript images (MIML) and used semantic web technology on some of the MASTER records made there (see http://www.raqoon.com/).

Although much has been done, much still remains to be done. The manuscripts catalogued thus far have been predominantly West-Norse (Icelandic and Norwegian), and chiefly literary; more experience is needed with Danish and Swedish manuscripts and manuscripts in Latin, and on other types of primary sources, such as diplomas.

*CHLT*

Another project in which the institute has become involved is CHLT (Cultural Heritage Language Technologies), a collaborative project involving other institutions in Europe and the United States: Department of English, University of Missouri at Kansas City (USA), The Perseus Project, Tufts University (USA), Department of Scandinavian Studies, University of California at Los Angeles (USA), The Newton Project, Imperial College, London (UK), Classics Department, Cambridge University (UK), Istituto di Linguistica Computazionale, Pisa (IT) and the Max Planck Institut, Berlin (DE). Funding for the project is provided by the National Science Foundation in America and the European Union International Digital Library Collaborative Research Programme. The project period runs from 1 June 2002 to 31 May 2005. The project has three major goals: first, to adapt discoveries from the field of computational linguistics and information retrieval and visualization in ways that are specifically designed to help students and scholars in the humanities advance their work; second, to establish an international framework with open standards for the long-term preservation of data, the sharing of metadata, and interoperability between affiliated digital libraries; and finally, to lower the barriers to reading Greek, Latin and Old Norse texts in their original languages (for more information see the project website: http://www.chlt.org).

The principal role of the Arnamagnæan Institute in the project is the provision of electronic texts, while our American partner, the Scandinavian Department at UCLA, handles the processing of these texts, in particular the development of a morphological analyser for Old Norse. This work has been carried out by a team of very capable students, under the direction of myself in Copenhagen and Prof. Timothy Tangherlini in California. It was decided to use eight of the Fornaldars Norlanda or mythical-heroic sagas, which deal with the early history of Scandinavia, as a test corpus, basing our texts each on a single manuscript, normally the oldest but in any case the one deemed to be the best. All the texts are marked up using TEI-conformant XML. The transcriptions have generally been made on the basis of a printed edition, but as few of the extant editions reproduce the text of the original manuscripts as diplomatically as we wanted, a good deal of "un-normalisation" has been necessary. At the same time, a fully normalised form of every word is added to the mark-up for search and processing proposes.

In brief, the transcription conventions we have employed are as follows: The text is transcribed exactly as it is in the manuscript with respect to orthography and spacing between words. Variant forms of the same letter (allographs) are not distinguished, apart from small capitals, used to denote geminates (principally N and R, but potentially also D, G, M, S and T), high and round s, ordinary and round r (r-rotunda), ordinary and insular forms of f and v, ordinary and uncial forms of d, e, m and t, all of which are represented using entity references. Only ligatures with an independent phonemic value (a and e, double a etc.) are represented; ligatures which are the result of graphic economy are treated as two separate characters (high s + t, for example). Abbreviations are expanded in accordance with the normal spelling of the scribe in question, using `<expan>` to indicate supplied letters, and the means by which the abbreviation is achieved, i.e. the sign or tittle used expressed as an entity reference, is given as the value of the abbr attribute. Abbreviation by suspension is distinguished from abbreviation by other means (contraction, supraliner symbol etc.) by means of the type attribute so that these may be processed differently. Letters or words assumed to have been inadvertently omitted by the scribe (which in a printed edition would normally be placed in angle brackets) are supplied and tagged using `<supplied reason="omitted">`, while `<supplied reason="illegible">` is used to indicate letters now unreadable but assumed originally to have been in the manuscript (which in a printed edition are normally placed in square brackets). Where necessary to the sense, emendations and alterations are made to the text; obvious misspellings, for example, are corrected using `<corr>`, with the original reading given as the value of the sic attribute. Additions and deletions made in the manuscript by the scribe or in another hand are indicated with the `<add>` and `<del>` elements. Line-, column- and page-boundaries are indicated using the empty milestone tags `<lb/>`, `<cb/>` and `<pb/>`, giving a number for each as the value of the n attribute. Large structural

divisions in the text, i.e. chapters, are tagged using  and given a number. Chapter headings are tagged using <head>, and the nature of the <head>, i.e. whether it is found in the manuscript itself or supplied by an editor, indicated in the value of the type attribute. The many verses in the text are tagged using <lg> (line-group) for stanzas and <l> (line) for individual lines. Owing to the prosimetric form of much saga literature, verses normally occur within prose paragraphs; this has necessitated changing the DTD in order to allow <lg> to appear directly within <p>. Finally, each word in the text is placed inside an <orig> element, and the normalised form is given as the value of the reg attribute. Compound words written separately in the manuscript should be grouped together within a single set of <orig> tags, while in the opposite situation, where for example a preposition and its object are written as a single word, the two parts are treated as separate words, each placed within a set of <orig> tags, but with no space between them. Marks of punctuation are placed outside the <orig> tags. Although relatively simple, this mark-up allows for (at least) three separate views of the text – strictly diplomatic, retaining the line-breaks, variant letter forms, unexpanded abbreviations and so on of the original, semi-diplomatic or semi-normalised, where the abbreviations have been expanded and any obvious errors have been corrected, and normalised, where spelling, capitalisation, word division and so on have all been regularised – through the use of multiple style-sheets, allowing the user to decide which view he or she prefers (and the ability to toggle between them). Clicking on a word opens a window providing a translation and grammatical and morphological information, which is extremely helpful to students. We hope also to provide links to digital images of the manuscripts themselves, at least on a page-by-page, but possibly on a line-by-line basis.

## Author Information

M. J. Driscoll – Arnamagnæan Institute, Copenhagen; e-mail: mjd@hum.ku.dk

# THE LATEST PRAGUE CONTRIBUTIONS TO WRITTEN CULTURAL HERITAGE PROCESSING [1]

Kiril Ribarov

*Abstract: This work presents a software package ACT (Annotated Corpora of Text) for lexical and corpus processing of European written cultural sources (currently used for processing of mediaeval Slavonic manuscripts). I use ACT as a contribution towards a contextual and intelligent heritage Information Technology framework. The software is suitable for capturing characteristics of old written sources including rich language variability on word and sentential level. It is not the word-form, but its understandings/interpretations that become central processing units, which can be assigned morphology distinctions, head-words (including recensional), translation equivalents; these interpretations can be joined in multi-word units or assigned correlation to other sources. The whole annotation process is automated and individual sorting orders and morphology tags structures can easily be defined. ACT incorporates modules for: complex searches on one or more sources, creation of various ready-to-use documents, web text and image access, incorporation of lexical card-files into a corpus, and text-from-card-files reconstruction.*

*Keywords: annotation, Old-Church Slavonic, lexical processing, cultural heritage*

[1] The following text has been originally published in the Proceedings of the Language Recourses and Evaluation Conference held in Lisbon, Portugal, 2004, under the title of "Towards Intelligent Written Cultural Heritage Processing - Lexical processing". I present here a revised contribution of the aforementioned paper and I add here the latest efforts done in the Center for Computational Linguistic in Prague in the field under discussion.