

**National Council on Orientalist
Library Resources Information
Technology Sub-committee**

Electronic Manuscript Catalogues

A report on current trends in electronic manuscript cataloguing and their implications for Oriental manuscript collections

Introduction

Manuscripts : the last stronghold of the manual catalogue record

Manuscripts have so far remained outside the electronic catalogue revolution that has swept the world of the printed book. Their difficult and complex nature and the fact that a printed manuscript catalogue is a piece of scholarly work for which the author expects to receive academic recognition has led to an idiosyncratic printed corpus of literature with few standard reference points. As highlighted in Kim Plofker and David Pingree's article on electronic cataloguing of non-Western manuscripts¹, oriental manuscripts offer particular challenges.

"The task of descriptive cataloguing itself, irrespective of the format of the finished product, exhibits a new set of problems to those cataloguers who venture outside the European textual tradition. In the first place, the greater chronological extent of the chirographic tradition in most of Asia, as compared to that in most of Europe (where scribal copying of texts began later and was sooner replaced by printing) means that there are simply more manuscripts from the non-Western world, by some orders of magnitude. Related to this is the comparative lack of bibliographic control for non-Western manuscripts, where the relation among author, text, and manuscript is generally much more obscure than for European ones. Thus any catalogue of such manuscripts must undertake to be also a catalogue of the texts themselves, rather than merely concentrating on the physical characteristics of manuscript instances of known works.

The same uncertainty permeates almost all other aspects of Asian manuscriptology. Even the subject classifications of known works are not firmly established in all their divisions: although a sufficiently minute classification scheme may exist in the work's indigenous tradition, such schemes are in many cases not yet usefully integrated into Western bibliography. The features of the physical manuscripts are no more tractable than those of their more abstract contents. Since it is neither feasible nor sensible to refrain from cataloguing non-Western manuscripts until all these varied issues of bibliographic and codicological control have been resolved (and indeed, in all probability it is only the knowledge gained through cataloguing attempts that will make it possible to resolve them), the author of such a catalogue must for the present rest satisfied with ensuring that its structure and presentation convey as much information as possible, without giving misleading impressions of certainty where none exists."

The success of standard bibliographic descriptions for printed books and increased ease of access which union databases such as RLIN and OCLC have brought to the scholarly community, however, have meant pre-eminence of the printed manuscript catalogue is now being called into question by funding bodies and scholars.

- Top-level managers of institutions and funding bodies have been influenced by the idea that electronic is best (preferably with digital images) and are increasingly only prepared to fund electronic resources

¹ **A Text-Based Approach to Electronic Cataloguing of Non-Western Manuscripts**
[Web page last accessed 9/11/01]

- A new generation of computer literate academics want to use and participate in the creation of electronic manuscript archives

Both trends put pressure on libraries and archives. Young academics find it easy to obtain funding for individual research projects involving electronic manuscript descriptions and digital images from funding bodies who have not always thought about strategic issues such as the use of common technical standards.

The report that follows shows how complex many of the issues are and how far we still are from reaching a set of common standards for cataloguing oriental manuscripts but there is a real danger that if the library world does not make progress in agreeing on standards it faces a future in which it is called upon to manage an nightmare of heritage data from incompatible electronic archives.

Electronic Manuscript Cataloguing Standards

MARC versus the XML/SGML approach

MARC : history

An early approach was to investigate whether the existing MARC standard could be used for manuscript records. MARC was considered well placed as a candidate for the task due to its overwhelming success with book records but as it stood the fields used were not wholly suited to use for manuscript records. They did not describe enough about the item, let alone its position within a collection and the complex hierarchies that are often involved in manuscript cataloguing. During the early 1980's the USMARC AMC (Archival Material Control) standard was developed. This widened the MARC format for use by the Archive community.

The original MARC format uses AACR2 cataloguing rules but AACR2 had not found favour with the manuscript cataloguing community because its chapter on manuscripts abandoned longstanding descriptive principles for archives. A new descriptive standard was therefore developed in the US to complement the MARC AMC standard; APPM Archives, Personal Papers and Manuscripts developed by Steven L. Henson at the Library of Congress. This standard was subject to scrutiny by the Library of Congress, where Henson worked as Senior Manuscript Cataloguer, and the American archival community. The first edition was released in 1983 and revised for publication in 1989.

MARC : the strengths

The obvious advantage offered by the MARC standard is the ease with which printed book records and manuscript records could be integrated. Libraries already have the mechanisms to exchange and display MARC records.

MARC : limitations

The principle limitation of the Archival Material Control MARC standard is that, for all its enhancements, the basic record length is limited and restricts how much information can be recorded about collections or individual items. This would considerably hamper the retrospective conversion of printed manuscript catalogues, which usually describe individual items and collections at considerable length. The restrictive length of MARC records may allow for the creation of useful finding aids but they could not be used as research tools in the same way as printed catalogue entries.

Although within MARC AMC empty fields have been adapted to demonstrate hierarchical relationships between items and collections the scope to describe complex relations is still fairly limited.

MARC software, because the standard is confined to libraries, is relatively expensive. This is not an issue for institutions who are already using MARC for their printed books but would represent a significant investment for small archives.

The XML/SGML approach : history

SGML is Standardised General Mark-up Language. SGML is an ISO standard (8879) which was developed in the early 1980's. It allows for the definition of content and structure for different types of electronic documents. It can be thought of as a language that is used to create other languages. SGML, for example, is the language that is used to define Hypertext Markup Language (HTML). HTML has served the World Wide Web very well and as has grown over the years to accommodate advances in the way people organise and use information. An example of this is the <TABLE> tag, which was not part of original HTML, but was added later because it is a very useful way to organise information on a Web page.

There is, however, a fundamental problem with HTML as a way to organise complex data.

- **HTML does not mark up content, only appearance**
- **HTML style is not rigidly enforced**

Since HTML was designed to format data for display, the majority of the tags are designed for appearances, not for data. For example:-

```
<H3>Paul Skilling</H3>
<I>Pali manuscripts</I>
<B>March 1, 2000</B>
<CENTER>
Anguttaranikaya
</CENTER>
```

In the small chunk of HTML above there is a person's name as a level 3 Head, then the words Pali manuscripts in italics, a date in bold and Anguttaranikaya centred on the page. It is not clear from the HTML, however, what the information means. It could be a list of a Pali manuscripts Paul Skilling bought in March 2000, it could be a list of Pali manuscripts that Paul Skilling wants or an article by Paul Skilling on the Anguttaranikaya. There is no way of telling from the HTML.

HTML style is not rigidly enforced and today's browsers are very liberal when it comes to interpreting HTML., but while Netscape and Explorer might both forgive mistakes in HTML, they might each forgive them in a different way. Incorrect HTML could simply result in a poorly displayed page in one browser but a complete loss of some elements of data when displayed in another browser.

SGML, which is rigidly enforced, could be used for Web based document management, but while SGML is a very powerful language, its draws its power from its complexity. Using SGML in conjunction with the Web is a very complicated undertaking and has never really caught on in the mainstream. It became apparent to many members of the Web community that there was a need to simplify SGML to make it more accessible for Web users. Thus, the SGML Editorial Review Board became the XML Working Group, chartered with creating the XML Recommendation for the World Wide Web Consortium (W3C).

The goal behind creating XML was to remove some of the lesser known features SGML and to define more clearly some of the syntax and structures of SGML, making easier for users to learn XML quickly, and eliminating confusing SGML applications. XML documents have tags which define data elements and are validated against a set of rules for their creation. This set of rules for XML is called Document Type Definition (DTD). The data content of the earlier example of information in HTML becomes clear if tagged according to XML.

```
<AUTHOR>Paul Skilling</AUTHOR>
<ARTICLE>Pali manuscripts</ARTICLE>
<PUBLISHED>March 1, 2000</PUBLISHED>
<SUBJECT>Anguttaranikaya</SUBJECT>
```

Paul Skilling is the author of an article *Pali manuscripts* published on March 1, 2000 and the subject of the article is the Anguttaranikaya.

The XML/SGML approach : strengths

In the first place XML files are not limited in length, unlike MARC records. The varying lengths of finding aids and item records do not, therefore, present an obstacle to record production. This also means that institutions can successfully produce their printed finding aids from XML records, as no data has to be omitted.

XML can support a potentially infinite number of collection levels. Multiple and complex relationships between items and collections can be easily represented within an XML file, thus not compromising the quality of an institution's records.

The investment in XML resources including effective search tools is not restricted (unlike MARC) to the library and archive communities. XML is being universally supported because it offers to a wide variety of differing markets a powerful tool with which many new web based applications can be generated. Although XML is a sufficiently new development that support is at present patchy, the library and archive sector will in future be able to reap the benefits of the significant funding that is likely to be put into developing applications for the diverse communities using XML.

The XML/SGML approach : limitations

For manuscript XML DTDs to work for Oriental materials three elements are needed, all of which are in early stages of development.

- XML editors and parsers (the validation programmes which check XML data against the DTD) which can handle Unicode in a user-friendly way. Many editors at the moment can only display Unicode characters as codes. This makes proof reading within the editor extremely difficult. Parsers, at present, frequently read valid Unicode characters as errors causing correctly tagged documents to fail the validation process.

- Web browsers which can handle XML documents and Unicode characters. Web browsers commonly in use at the moment vary in their ability to handle XML and/or Unicode.
 - Search Engines which can perform complex searches using XML tags. Search Engines which can exploit the detailed tags that can be put in DTD's are still in the early stages of development.
-

Three XML/SGML based manuscript description standards

EAD (Encoded Archival Description), MASTER (Manuscript Access through Standards for Electronic Records) and ACSAM (American Committee for South Asian Manuscripts)

EAD : history

Encoded Archival Description is a project which has developed a DTD for the purpose of generating online finding aids at the University of California at. It uses the ISAD(G) standard for archival description as developed by the International Council of Archives and SGML/XML. The Library of Congress and the Society of American Archivists jointly maintain it. EAD developed from a finding aid research project at Berkeley and really took off in 1995.

The key decisions as to what should be included and which elements should be represented were made by a panel of archivists as opposed to SGML experts and the decision making progress is recorded in depth within the Ann Arbor Accords. The EAD DTD version 1.0 was released in February 1996. Progress has been made with the DTD being constantly monitored and revised as issues have been raised.

There has been widespread interest in the US, where a large number of institutions have used the EAD DTD to produce electronic records documenting their collections The Research Libraries Group (RLG) has over 600,000 records describing its archival resources and maintains a series of resources on their website for those interested in using EAD². In the UK, the Public Record Office (PRO) has adopted EAD as part of the Access to Archives initiative and has recently gone live with a finding aid. The PRO has been keen to promote EAD and offers a series of resources on its "Access to Archives" website³. EAD is also being used for the India Office Records at the British Library. The newly created HE sector

² **RLG EAD Support Site** [Web page last accessed 5/11/01] <http://www.rlg.org/rlgead/>

³ **Public Record Office A2A: Access to Archives: the English strand of the UK archives network** [Web page last accessed 5/10/01] <http://www.pro.gov.uk/archives/a2a/>

Archives Hub⁴, which aims to become a major national gateway to archive collections held in UK universities and colleges is using the EAD DTD for its records.

EAD : strengths

EAD has had a number of years development and has the support of such major institutions as RLG, the Library of Congress and the PRO. The UK based Archives Hub has financial backing from the Joint Information Systems Committee (JISC). This support indicates that EAD's long term future is secure.

The adoption of EAD by a number of prominent institutions has also driven the development of tools to make it easier to create and to search EAD records. The Archives Hub website offers searches for Subject, Place Name, Date, Title, Genre/Form and Full Text. It also offers guidance and templates for data creation. The European project MALVINE (Manuscripts and letter via integrated networks in Europe)⁵ is also developing search facilities for navigating EAD records and expects to have them fully operational by the end of 2001.

EAD : limitations

EAD is, first and foremost an archival description tool, reliant on ISAD(G) guidelines for producing its records. While it would be possible to produce finding aids for oriental manuscripts using EAD any more detailed description would probably need additional tags, which would be unlikely to be accepted for inclusion in ISAD(G) revisions by the International Council of Archives without considerable lobbying from the oriental library community. The other alternative would be to stretch the interpretation of existing tags to cover oriental manuscript needs, but this would seem to defeat the purpose of XML DTD's, which are, after all, a means to provide methods of description and organisation precisely tailored to the material in hand.

The archival community has tended to be resistant to the idea of national/international Authority Files. Authority files for EAD records appear to have been left to individual projects, which will cause problems as union resources are developed.

EAD is a top-level down approach moving from collection level description to the particular item. Printed oriental manuscript catalogues have tended to use the bottom-level up approach, concentrating on detailed descriptions of individual items. This could present considerable problems if printed catalogues were to be converted into EAD records.

MASTER : history

MASTER is a European Union funded project to create a single on-line catalogue of medieval manuscripts in European libraries⁶. MASTER started in January 1999 and finished in 2001 and was led by De Montfort University. The specific goals of the project were to develop a single standard for computer-readable descriptions of manuscripts, create software for making these records and test the standard and the software on descriptions of manuscripts. Over 800 of the records created have been mounted in online prototype networked catalogues at Leicester and Oxford (links available on the MASTER webpage). The MASTER project has strong links with a similar project in the US called

⁴ **Higher Education Archives Hub** [Web page last accessed 5/11/01]
<http://www.archiveshub.ac.uk/index.html>

⁵ **MALVINE homepage** [Web page last accessed 6/11/01]
<http://www.malvine.org/malvine/eng/>

⁶ **MASTER homepage** [Web page last accessed 6/11/01]
<http://www.cta.dmu.ac.uk/projects/master/>

Electronic Access to Medieval Manuscripts (EAMMS)⁷. EAMMS has been looking at the possibility of using MARC and SGML. for describing Medieval manuscripts.

Partners in the project include the Royal Library, the Hague; the Arnamagnaeen Institute, Copenhagen; L'Institut de recherche et d'histoire des textes, Paris; the National Library of the Czech Republic, Prague; the University of Oxford; The Bildarchiv Foto Marburg, Germany, and IBM UK are also partners in MASTER. In addition, several other major libraries and institutions are associated partners in the project: the Stofnun Árna Magnússonar á Íslandi, Reykjavík; the Universitetsbiblioteket, Lund, Sweden; the Narodna Biblioteka Sv Kiril i Metodij and Institut po Matematika i Informatika, Sofia, Bulgaria; The Perdita Project, Nottingham Trent University, UK; the National Library of Wales, Aberystwyth, UK; and Lietuvos Martyno Mazvydo nacionaline biblioteka, Vilnius, Lithuania. Other libraries which have participated in MASTER include the British Library, the Vatican Library, the Biblioteca Ambrosiana, and the Bodleian Library, Oxford. The project also has strong links to several North American manuscript projects and to the international Text Encoding Initiative, through a TEI workgroup on manuscript descriptions.

MASTER : strengths

This is an item based standard, which offers the scope for both basic and highly detailed description of manuscripts. The large quantity of tags, which have been developed in consultation with manuscript specialists make this an extremely flexible tool. In addition to the manuscript description, it offers the opportunity to create list surrogates of the manuscript as part of the item level description allowing considerable scope for libraries to provide information about available microfilms, photographs and digitised images.

A union catalogue exists to which MASTER records can be contributed and free editing and parsing software has been created to help people in the creation of records.

MASTER offers the opportunity for oriental and western manuscripts to be catalogued using the same standard, which could enhance funding opportunities.

MASTER has tackled the issue of Authority Files and adheres to the Library of Congress Authority file for names, places and uniform titles. Where these fail, MASTER participants have created their own authority list⁸ and specific advice is given for constructing entries using AACR2.

MASTER : limitations

The MASTER project officially came to a close at the end of June 2001 and as yet no further announcement has been made regarding the future of the project and continued funding. The future of this DTD is therefore by no means assured.

MASTER has been driven by the needs of the western manuscript tradition and some additions would need to be made if it were to be adopted for oriental manuscripts. For example, the MASTER calendar attribute has recommended values in the reference documentation of Gregorian, Julian, Roman, Mosaic, Revolutionary and Islamic. These values would need to be expanded for Indic manuscripts to accommodate all the Indian eras and tags would also be needed for dates expressed in terms of phases of the moon and religious festivals etc. Requests to include tags to accommodate the fact that the relation between the author, text and the manuscript is generally much more obscure than for

⁷ Hill Monastic Manuscript Library *Electronic Access to Medieval Manuscripts* [Web page last accessed 6/11/01]

⁸ MASTER : Name Authority Resources [Web page last accessed 6/11/01]
http://www.bodley.ox.ac.uk/master/name_authority/name_authority.html

Western manuscripts may, however, be resisted on the grounds that such additions would make the DTD overly complex. The development team have indicated that they are willing to consider suggestions from the oriental community and, as the standard is likely to be adopted for a forthcoming catalogue of Jain manuscripts involving the Wellcome Library and the Bodleian, there will be opportunity to see how easy it is to influence the design of the DTD.

The MASTER project provides access to the Library of Congress Authority File through the Library of Congress Name Authority File at DRA.. The display strips out all diacritical markings so it is not possible to use it to authenticate oriental authors, uniform titles or places.

ACSAM : history

ACSAM (The American Committee for South Asian Manuscripts) was founded in 1995 under the auspices of the American Oriental Society to promote the preservation and use of manuscripts of South Asian origin in North American Collections. ACSAM intends to achieve this by

- identifying such manuscripts in public institutions and in private hands
- cataloguing manuscripts through training and deploying Graduate Students to effect this task under guidance and with the assistance of members of the Committee
- encouraging owners of manuscripts to conserve those in need of attention
- making a digital version of the catalogue and digital images of every leaf of every manuscript possible, which will be freely available to scholars throughout the world via the World Wide Web

ACSAM estimates that full implementation of this project will take between fifteen and twenty years. In July 2000 ACSAM launched a prototype online manuscript catalogue in collaboration with Brown's Scholarly Technology Group⁹. The prototype catalogue contains over 300 manuscript descriptions, mostly Sanskrit, but with some Arabic entries as well. The Scholarly Technology Group developed the DTD for the descriptive bibliography of South Asian manuscripts and a prototype Web delivery system for the data. Ongoing work on the project includes refining the DTD and experimenting with the delivery of accented and non-roman fonts on the Web and multilingual text retrieval.

ACSAM : strengths

ACSAM has the backing of the American Oriental Society, which provides some guarantee of its future existence. If the Union catalogue achieves its goal of entries for several thousand South And West Asian Manuscripts now held in North American collections, the underlying DTD will be seen as the natural standard for manuscript description by the many young scholars who will have been trained in its use as part of the programme.

The ACSAM DTD has been designed specifically with oriental script manuscripts in mind and solving the issues of delivery of oriental diacritical markings over the web and multilingual text retrieval is high on its list of priorities. The committee also places a high priority on developing a search engine that will, in addition to searching for free text and/or element type, or structural characteristic will employ various "fuzzy matching" techniques to optimise the usefulness of searches, whose typical ambiguities will include variant forms or vocalizations of names, comparisons of precise dates or chronological ranges with vague or unknown ones, identification of well known texts by the European version(s) of their titles.

⁹ **ACSAM Test Catalogue** [Web page last accessed 7/11/01]
<http://mama.stg.brown.edu:1084/dynaweb/acsam-test>

The ACSAM project has seen the need for development of authority files and the test catalogue demonstrates two authority files with brief descriptions of names of people and the geographical locations mentioned in connection with the manuscripts. They are listed in alphabetical order (according to the relevant alphabet). Each description contains links to the catalogue entries for other places and persons that it mentions and is shown with indications of the diacritical markings.

ACSAM : limitations

The ACSAM DTD is designed for the manuscripts of South and West Asia, which are primarily in Sanskrit, Arabic and Persian. It is not yet clear whether there will be other sub-committees of the American Oriental Society working on other oriental language manuscript traditions.

The project website does not indicate what progress has been made since the launch of the test catalogue in July 2000 and whether there is interest in sharing the DTD with oriental libraries in the UK. The Chair of the NCOLR IT sub-committee has tried to contact the STG lead Kim Plofkar but has so far not received a response. Conversation with the South Asian bibliographer of the University of Chicago has indicated that ACSAM has failed in two major funding bids to develop their work further and that the future of their XML catalogue is therefore extremely uncertain.

Conclusions

XML/SGML Solutions Gaining Ground

It would seem that the main focus of the archive and manuscript community is on developing descriptive standards based on XML/SGML solutions rather than MARC.

If the UK oriental library community follows the lead of these initiatives, there are at present three DTD's developed for use with manuscript and archival materials. There are two factors which really need to be considered before adopting a particular DTD.

- The DTD has to enable the material to be described as accurately as possible
- The DTD has to have a significant user group to guarantee its long term survival and the development of web based catalogues

It is possible to map elements from one DTD to another so they are by no means mutually exclusive but mapping does require software development. It is also possible to use a field in MARC to link to an XML description (this approach was being explored by Columbia University until they lost their source of funding for the project). Again such interlinking requires some software development.

EAD would seem to be the most appropriate DTD for collection level descriptions and finding aids with the MASTER and the ACSAM DTDs being suited to detailed descriptions of individual manuscript items. Until further information is received from the development team at Brown University, it is not clear whether there is any relationship between the MASTER DTD and the ACSAM DTD.

The main features outlined in the body of this report have been summarised in tabular form in the final section.

Summary of EAD, MASTER and ACSAM features

	EAD	MASTER	ACSAM
Started	1995	1999	1995
Descriptive Standard	ISAD(G)	Any	Unknown
Location	USA	Europe	USA
Agency responsible for DTD	Library of Congress /Society of American Archivists	De Montfort University	American Committee for South Asian manuscripts / Scholarly Technology Group, Brown University
Funding	Various US sources	EEC	Unknown
Status	Ongoing	Unknown	Unknown
Documentation available online?	EAD website	MASTER website	Not available online
Emphasis	Collection Level - Archival Resources	Item Level - Western Manuscripts	Item Level - South Asian Manuscripts
Catalogues available online?	Research Libraries Group / Public Record Office	De Montfort University / Oxford	Brown University
UK Users	Public Record Office / Higher Education Archives Hub / India Office Records	National Library of Wales / The Perdita Project, Nottingham Trent University ¹⁰	None

¹⁰ The Wellcome Library and the Bodleian may be using MASTER for a Jain manuscript catalogue to be sponsored by the Institute of Jainology and training courses for those involved in the project will be taking place early in 2002