Rendering complex descriptions: the use of EAD and TEI for the description of manuscripts and old books in France

Florent Palluault
French Ministry of Culture and Communication
Books and Reading Division
Paris, France

Meeting: 212 — *Cataloguing standards and special collections* — Rare Books and Manuscripts

Abstract:

The impending use of a specifically designed XML format combining MARC and TEI for the computerization of the French regional catalogues of incunabula prompts an assessment of the decade-long use of XML/EAD for the description of manuscripts and archives. After a consultation of national experts, EAD was selected for three major conversion projects (the catalogue of manuscripts held in French public libraries, the directory of 20th century literary manuscripts, and the national library's catalogues of manuscripts). A set of guidelines for the description of manuscripts and archives (DeMArch) and a best practice guide for EAD encoding, to be published shortly, will help librarians update, amend, and augment this mass of EAD data. While the national library and several academic libraries have been using EAD regularly for a few years, public libraries are slower to adopt the format. With EAD being seen as relevant for any set of hierarchical documents, the standard is spreading beyond manuscripts and archives to mixed collections of documents, as well as collections of photographs and audiovisual documents.

Librarians have acquired experience using EAD which will prove useful in the adoption of MARC-TEI. This XML format was designed to guarantee an accurate rendering of the lengthy and detailed records from the most recent volumes of the catalogues of incunabula. The records' global MarcXchange structure will include TEI segments for text transcriptions and some copy-specific data such as colophons, rubrication, and heraldry. Unlike the traditional MARC, this XML format will also be able to handle long citations, formatting, and special characters. MARC-TEI will also ensure interoperability with similar catalogues, especially with ISTC which is contributing to the project by providing bibliographic data; it will also be compatible with the descriptions of bindings, which the national library has started describing in TEI.

The computerization of the French union catalogue of incunabula (*Catalogues régionaux des incunables des bibliothèques publiques de France*) is scheduled to enter its execution phase later

this year. This project will implement a specifically designed XML format that mixes elements from both MARC and TEI (Text Encoding Initiative). XML formats have multiplied in libraries over the past ten years. This trend arose from the need for technical interoperability with existing data (exchange formats such as MARCXML and MarcXchange) and from the need for concise descriptive metadata (DublinCore, MODS, MADS) in the context of digital libraries. Only EAD (Encoded Archival Description) has been used extensively to catalogue special collections, and more precisely to encode descriptive records of manuscripts and archival documents. Therefore, the use of MARC-TEI for the description of incunabula in the near future justifies an analysis of the past and present use of EAD. Such an analysis may help anticipate any obstacles that might prevent endorsement of MARC-TEI.

France is a highly centralized country and, logically, the adoption of EAD owes more to a decision by several national agencies than to a gradual evolution of professional practices. In the early 2000s the ministry of Culture decided to computerize the union catalogue of manuscripts. At the time, the 116-volume *Catalogue général des manuscrits des bibliothèques publiques de France* (CGM), published from 1849 to 1993, comprised a total of 182,000 records of manuscripts and archival material. They covered the collections of 519 public and academic libraries, archival repositories, and various other institutions. The stirring committee, comprising representatives from the Ministry of Culture, the Ministry of Higher Education, the French national library (BnF) and the Bibliographic Agency for Higher Education (ABES), commissioned a consulting agency to provide a preliminary study on the best suited formats. MARC, EAD, and various other solutions were analysed, and the stirring committee followed the consulting agency's recommendations by selecting EAD in the Spring of 2002. Around that same period, the BnF initiated the computerization of the Manuscripts Department's numerous legacy finding aids and catalogues and also chose EAD as its target format.

In 2002, selecting EAD as a cataloguing format for French libraries was not a given. Until then, this standard had not gone further than a few experimentations and was perceived as an import from archivists’ practices which was largely foreign to the library world. Several points of view on the cataloguing of manuscripts and archives coexisted. Some librarians advocated concise descriptions created in a spread sheet, a database, or even in the MARC structure familiar to librarians. However, even if short-title catalogues allow libraries to catalogue quickly great numbers of documents, they rarely satisfy researchers and are not suited to the computerization of legacy finding-aids' lengthy descriptions. It was imperative to preserve the wealth of information contained in printed records: beyond the necessary elements of bibliographic information (responsibility statement, title, date of creation or publication) the value of the description of ancient documents often lies in the details that reveal their historical significance: uniqueness, rarity, origination, provenance, physical characteristics, binding or container. Through time, the way all these pieces of information are presented in printed catalogue records has gradually settled. Therefore, some manuscript cataloguers wanted to keep using word processors, which allow for a satisfying layout of printed or PDF publication, even if the absence of semantic structure precluded any efficient search through the data.

The CGM preliminary study determined that EAD guaranteed the interoperability on the web inherent to XML formats, and more importantly a great number of adequate semantic elements and a structure well adapted to the diversity of existing catalogues, in particular those containing records covering multiple hierarchical levels. The professional consensus in favour of EAD prevented the divergent choices seen in the cataloguing of printed documents in the 1990s when university libraries and later public libraries opted for UNIMARC instead of the national library's INTERMARC.

The computerization of the CGM, which represents the largest retrospective conversion of a printed union catalogue in France, was completed in April 2008. The computerization of the BnF Manuscripts Department catalogues was more complex, due to the diversity of the finding aids and the multiple languages and alphabets used ; it is now almost complete. In addition to these two main endeavours, a third, more modest catalogue, the directory of 20th century French literary manuscripts (*PALME*), was converted from INTERMARC to EAD in 2007. These projects, initiated by government agencies and by the BnF, quickly created a critical mass of EAD data: an estimated two-thirds of the manuscripts and archival material held in French libraries is today described in EAD. The records are distributed among three online catalogues : *BnF Archives et manuscrits* which describes the collections kept at the national library; the catalogue of archives and manuscripts of the Higher Education institutions (*Calames*), and the manuscripts subdomain of the French union catalogue (*Catalogue collectif de France*, CCFr) which will eventually provide a common gateway to all EAD data from French libraries.

The preliminary studies of EAD implementation in libraries and the debates over the format's characteristics at workshops attended by both librarians and archivists quickly highlighted the absence of common rules for the cataloguing of manuscripts and archival documents held in libraries. Whereas standards had emerged quite early for the description of printed books, the standardization of catalogues of manuscripts, for which data exchange is obviously less crucial, had gone no further than formalizing local practices. The General International Standard Archival Description (ISAD(G)), geared towards archival repositories, was only rarely implemented in French libraries. This normative gap was filled in 2010, when the French standardization agency published DeMArch [*Description des Manuscrits et fonds d'Archives modernes et contemporains en bibliothèque* = Description of modern and contemporary manuscripts and archival fonds in libraries]. Although DeMArch is a set of guidelines and not an official standard, it holds the full scientific value of a standard. DeMArch was intended to be a subset of ISAD(G); it used DACS (*Describing Archives: a Content Standard*) as a model and adapted it to French libraries. DeMArch states the main principles of archival description and explains how to express the various pieces of information (identification, context, content and structure, restrictions of access and use, acquisition, appraisal, etc.). DeMArch is intended to be a set of descriptive rules independent from any particular computer format, and therefore offers examples of records taken from various types of finding aids: EAD, MARC, Dublin Core and relational databases.

The DeMArch guidelines are quite recent and are not yet widely in use in libraries. However, they are bound to spread further with the impending publication of the best practice guide for EAD in libraries (*Guide des bonnes pratiques de l'EAD en bibliothèque*), which used DeMArch as a model. This guide endeavours to compensate EAD's inherent flexibility. The heterogeneous implementation of EAD has many causes: a great number of elements are available at each description level, recursion is allowed in several elements, the DTD offers little control over input while the tag library allows quite a degree of interpretation, and data is created and accessed through various modes of production and dissemination (indexing and publishing). Flexibility is useful because it allows EAD to cater for any type of existing record, but it also proves tricky in regular cataloguing, especially since XML editors, which are predominantly used in France, do not hinder cataloguers' creativity. In 2008, in order to ensure data consistency across union catalogues, a group of manuscripts experts was given the task to write this EAD best practice guide, which should get published by the end of 2012. Creating the guide took longer than expected as it became necessary to find answers to issues regarding description as well as encoding. The debates highlighted variations in practices between various libraries and consortia, and even among the departments of the national library. They also showed members of the group the adjustments necessary for the detailed EAD encoding of non-manuscript documents, such as audiovisual material and objects.

The best practice guide is targeted mainly at EAD cataloguers, but will also prove useful for the conversion of remaining printed catalogues. The guide offers general recommendations about encoding, as well as more detailed information about the use of some EAD elements, attributes and attribute values. It also endeavours to distance itself from specific production and publication software, but does not skirt issues such as indexing and publishing, the influence of cataloguing software over encoding, or libraries' institutional strategies. The *Guide* aims to rectify some early implementations of EAD imposed by the large CGM and BnF computerization projects which took some liberties with the format. It seeks a compromise between practices born of these main conversion projects and the freedom requested by cataloguers wishing to create lengthy and elaborate records. For example, whereas the CGM conversion used 66 out of the 146 EAD elements, and PALME only 41, the *Guide* allows 100 elements while giving precise information on their use.

The challenge facing many manuscript specialists for the next 10 years will be to make use of, amend and add on to the CGM and other online catalogues. Each library can obtain the EAD records corresponding to their collections from the French Union Catalogue Service (CCFr) or the Higher Education Bibliographic Agency. They can then update the records and add descriptions of any document acquired since the publication of the last printed volumes.

University and other higher education libraries can already update and publish EAD data on a common platform: the Calames application comprises both a discovery interface and a cataloguing module composed of an XML editor with add-ons that allow links to a common authority file. Gradually, the number of libraries using this software is increasing beyond the original cluster of institutions involved in the CGM conversion project.

Around 2005-2006, the French national library (BnF) developed its own piece of software to create EAD descriptions. Like Calames, PiXML is based upon an XML editor and manages links and data exchanges with the BnF authority files. For dissemination, the BnF selected Pleade, a software specifically designed for the publication of EAD finding aids. Currently, several departments of the BnF use these applications: the Manuscripts department, the Performing Arts department, the Arsenal library, and the Audiovisual department.

The implementation of EAD cataloguing in public libraries is a much slower process. Except in a few major city libraries, it faces many obstacles, whether they be technical (absence of adequate cataloguing and publication software), financial (high cost of computer developments), or human (lack of EAD-trained librarians). The Ministry of Culture could not commission an IT company to create a common cataloguing software for public libraries because it cannot impose any specific software solution upon local councils. In order to take advantage of the momentum provided by the CGM computerization, the Ministry and the Union Catalogue Service (CCFr) are looking at several options. In collaboration with the University of Caen they are producing a prototype for a free piece of software which would provide EAD cataloguing and an easy HTML output, so that libraries may publish the records of their manuscripts on their own website. Furthermore, they started experimenting on the CGM update at the regional level. In June 2012 a study was launched in the Champagne-Ardenne region to assess the needs of libraries and the means available: the results will help define the modus operandi by which the three main regional libraries and the local library cooperation association may help smaller libraries update their catalogues of manuscripts. In the meantime, the CCFr keeps making global corrections to the CGM data, provides both training in EAD cataloguing and expertise on record conversion, and if necessary offers financial aid for EAD description projects. The 2005-2006 survey of heritage collections in public libraries ordered by the Ministry of Culture as part of the Action Plan for Written Heritage (*Plan d'action pour le patrimoine écrit*) confirmed that about a fourth of manuscript collections in public libraries have yet to be properly described.

EAD training programmes keep emerging. The national library has been offering some EAD training for its own personnel for many years, and opened up to other libraries in 2008. The national school for information science and libraries (ENSSIB), which trains librarians and library curators, included EAD in its curriculum almost 10 years ago. Several smaller library schools also offer theoretical and practical EAD training as part of continuing professional education. Once the best practice guide for EAD is published, these various training programmes will benefit from reliable documentation specifically aimed at libraries.

An advanced EAD training session is also being planned for the end of 2012. It will help librarians who are already familiar with the standard take full advantage of XML functionalities. They will learn how to use XPath queries within EAD finding-aids and how to employ XSLT transformations for many uses: extract data from or add data to finding-aids; create new enriched finding-aids, collections analysis tools; create programmes to convert Word, Excel or MARC catalogues to EAD; create Dublin Core metadata, etc.

This technological know-how will become increasingly necessary as the realm of EAD keeps expanding. There is a logical consensus on the use of EAD for the description of archival fonds in France, and EAD is also the de facto standard for manuscripts. Some librarians, however, have criticised the use of EAD for single manuscripts, particularly medieval manuscripts. Indeed, a tree structure is less necessary for these documents than for archives, and EAD semantics are not as developed as those of TEI-P5, which is being used in several European countries for the description of medieval manuscripts. For example, the detailed physical description of a medieval manuscript requires the use of several EAD <physfacet> elements, each with a precise value for the type attribute, whereas TEI provides several independent elements to encode the same information, and greater flexibility in the transcription of incipits and explicits.

Beyond manuscripts and archives, EAD is relevant for any set of documents linked by hierarchical relations, whatever the nature of the documents may be. EAD allows to display easily the structure of fonds or mixed collections such as those of the Performing Arts Department at the national library, which often comprise manuscript scenarios, printed programmes, costumes, objects from the theatrical scenery, etc. EAD can also be applied to archives of writers composed of archival documents, manuscripts, and printed books, or to archives of architects that include project files, blueprints, drawings, and photographs.

Printed documents that are part of an archive may be described in EAD in a succinct manner and a link added towards a more thorough description in an independent MARC catalogue, since MARC is more adapted to the description of published documents than EAD is. EAD's advantages in describing hierarchical sets of documents also prompted the national library to catalogue part of its collection of discs and videos in EAD. Academic libraries also selected this standard for their large collections of photographs. Generally, MARC remains the standard for visual documents, but EAD is progressing and there is a plan for a national meeting to discuss whether a standard should be preferred over another.

With EAD gaining momentum in French libraries, specialists of the standard are keeping a close eye on its current evolution. The technical subcommittee in charge of preparing the EAD 2013 schema at the Society of American Archivists comprises a French representative. Until now, French institutions have almost exclusively been using the 2002 DTD version of the standard, with some characteristics specific to libraries (unnumbered components, shared lists of attribute values for some elements, use of UNIMARC or INTERMARC relator terms for the role attribute of person, family and corporate names). The switch to an XML/EAD schema will obviously entail some software update, but there is some specific concern about the underlying changes to the format.

Some of these evolutions are a logical follow-up of alterations operated since EAD 1.0, and do not call for any particular comment. French professionals, however, feel that the current direction towards a much more database-compatible format might be detrimental to the nature of EAD finding-aids, which are XML-encoded descriptive texts rather than a series of data fields. EAD 2013 will also be more closely compatible with EAC-CPF, which is EAD's counterpart for archival authority records. EAC-CPF is under experimentation at the national library but is not being used elsewhere so far, with libraries generally trying to exploit existing MARC authority files.

Whereas EAD now enters a phase of expansion, consolidation, refining of practices, and improvement of production and discovery tools, MARC-TEI will soon be implemented for the first time in a much narrower field: the cataloguing of incunabula. In 2009, the Ministry of Culture decided to computerize the union catalogue of incunabula (*Catalogues régionaux des incunables des bibliothèques publiques de France*, CRI). The 16 volumes published since 1979 cover about half of the copies kept in French libraries. Another 15 volumes are planned in the coming years. During the feasibility study by the *Centre d'Etudes Supérieures de la Renaissance* (Centre for Advanced Studies on Renaissance, CESR, University of Tours), a group of experts in old books and bibliographic formats tested the implementation of EAD, TEI, MARC, and of a model specifically designed for the project, on the description of incunabula. It appeared that the data structure which guaranteed both an accurate rendering of the lengthy and detailed records from the most recent volumes, and interoperability with similar catalogues required to use MARC and TEI simultaneously. This new format was expressed as an XML schema based upon the MarcXchange schema. The expert group included members of the commission appointed to revise the AFNOR Z44-074 standard for the cataloguing of old books (French counterpart to ISBD(A)), and the current evolution of this standard was therefore taken into account in the definition of MARC-TEI.

MARC-TEI was designed so that the overall structure of incunabula records is in MARC (MarcXchange) and is therefore entirely compatible with records of post-incunabula, 16th Century books and other old books. The TEI segments within some MARC subfields allow to include text transcriptions, and to describe precisely incipits, colophons, rubrication, or heraldry, with a precision that MARC cannot hope to match. Only some very specific TEI elements, which will bring some rich semantics that are lacking from MARC, will be used. These elements will be employed to create indexes that will be exploited by search criteria. The list of these TEI elements is not fixed yet, as it may vary according to the way TEI encoding will be carried out during the conversion process: the additional TEI tags may be placed in the data as each record is analysed and distributed into MARC fields or, more likely, after all the MARC data is input in the database. The XML context will also allow an adequate rendering of formatting (superscript characters or indices in signatures, italics) and rare symbols and characters. The citation of lines of text to differentiate variants requires to use special characters and abbreviation symbols which stem from the manuscript tradition and have not necessarily been validated by the international Unicode authorities. The CESR therefore plans to request, via the Medieval Unicode Font Initiative, that some character types necessary for the description of incunabula be included in the next version of the Unicode standard.

This computerization project and more generally the study of descriptive formats and their computer equivalent have provided an opportunity to look beyond the French borders. In November 2010, the CESR and the Ministry of Culture organised a workshop dedicated to the cataloguing and digitization of incunabula. Librarians from the Incunabula Short Title Catalogue (ISTC, British Library), the *Gesamtkatalog der Wiegendrucke* (GW, Berlin State Library), the Bavarian State Library, the Consortium of European Research Libraries (CERL), the National Library of France and several other French libraries were invited. An overview of the various cataloguing formats used by European partners strengthened our opinion that MARC-TEI was the right choice. The two main union catalogues implement an XML solution: ISTC uses MARCXML, which is well suited

to the needs of a short-title catalogue, while GW uses its own XML data structure but is also considering moving towards TEI.

This workshop also turned a budding collaboration with the British Library into a reality. John Goldfinch, head of the ISTC, greatly helped the CRI project by providing the CESR with raw data from the ISTC. This external data input will allow the CESR to group automatically records related to the same edition that appear in different volumes of the catalogue. The next step will consist in selecting the most precise and up-to-date record, and possibly in adding some information from other records. Any copy-specific data (library holding, shelfmark, binding, handwritten inscriptions, provenance) will then be added on to this new authoritative record. The regional partition of the catalogues will therefore disappear and be replaced with a unified national perspective. The computerized records of the *Catalogues régionaux des incunables* will thus create a national union catalogue of incunabula intended to host descriptions of any 15th Century printed document kept in French libraries, the national library included. The BnF's *Catalogue des Incunables de la Bibliothèque Nationale*, which is nearing completion, could be computerized using the same processes and software as the regional catalogues.

The computerization will also provide an opportunity for creating a cataloguing application specific to incunabula, to be used for catalogues currently in process and future catalogues. The CESR is considering using the Koha open source information library system for the data conversion. Koha's interface is more user-friendly than a regular XML editor and Koha will make it easy to manage the MARCXML data provided by the ISTC. During the feasability study, the CESR adapted Koha in order to include custom fields for copy-specific data (variants, rubrication, binding), and to add TEI segments easily. Systematic links to the BnF authority files for authors, printers and any identifiable former owner are also planned. The description of incunabula kept in regions not yet covered by the catalogues would consist in identifying the corresponding existing record in the database, and adding a copy with its localisation and copy-specific information. For incunabula editions not yet described in other catalogues, the Koha interface would provide the fields and functionalities necessary to create detailed descriptions compliant with the rules used in the last printed catalogues. In the future, the public would access the online incunabula catalogue via the French union catalogue where a specific search interface, similar to the Manuscripts interface, might be developed in order to exploit fully the complex MARC-TEI format (with search criteria such as provenance, physical characteristics, etc.).

TEI is also being used at the national library's Rare Books Department for the description of the most precious bindings. The descriptive model was devised at the national library but is intended for use by any French library. Fabienne Le Bars, who heads this programme, presented the TEI format used for bindings at the IFLA conference in Göteborg two years ago (http://www.ifla.org/files/hq/papers/ifla76/99-bars-fr.pdf). Data is distributed into five main blocks of information (document identification, binding description, document history, identification of persons, families and corporate entities linked to the document, and bibliographic references). The combination of the TEI elements selected for this description covers all the information necessary for the description of a binding, as well as links to digital images of these bindings. This set of descriptions proves useful in identifying the various types of bindings and offers a vital aid to trace provenance information.

Conclusion

EAD was implemented for the description of manuscripts and archival material in a few major steps, thanks to the completion of massive computerization projects. This standard has now found some measure of success at the national library and in the main academic libraries which hold heritage collections, and is starting to spread in public libraries too. EAD's success will become

more asserted with the improved cataloguing tools, the greater number of EAD training sessions, and the emergence of a network of users ready to share their experience. The best practice guide will dispel some of the worries which still trouble newcomers and will enable librarians to amend and add on to the mass of homogeneous data born of the CGM conversion which is still mostly untapped. Librarians will become more familiar with XML and will be able to focus their efforts on refining their use of the EAD structure and encoding, on using XML functionalities and on designing more relevant and efficient discovery tools.

It is still too early to predict the future of MARC-TEI, but three aspects justify a positive outlook: first, training librarians in using this format should be easier than EAD training. Usually, incunabula specialists already know MARC quite well, and the TEI additions will be limited to a few elements which will be defined and explained in a user guide. Furthermore, any XML experience acquired from the use of EAD will help libraries evolve towards an XML version of MARC. Finally, once the conversion of the CRI catalogues is complete, the cataloguing effort will be minimal: whereas there are still huge manuscripts and archives collections to be described, never-before described incunabula are rare discoveries. In the future, the cataloguing process will focus on adding new locations and copy-specific information to existing bibliographic records. The next challenge will consist in linking these records of incunabula, when relevant, to TEI descriptions of bindings.