The TEI-XML Architecture of *Ethiopian Manuscript Archives*: Respecting the Integrity of Primary Sources and Asserting Editorial Choices

Anaïs Wion, Centre National de la recherche scientifique, Institut des Mondes Africains

The paper illustrates the choices made by the *Ethiopian Manuscript Archives* collaborative project when defining the TEI-XML schema and architecture for the encoding of charters and similar documentary texts transmitted within Ethiopian manuscripts.

The *Ethiopian Manuscript Archives* (EMA) is a collaborative project carried out by historians and philologists working on manuscript documents produced by the Ethiopian Christian kingdom between the tenth and the twentieth centuries. It was developed at the Institut de Recherche et d'Histoire des Textes (IRHT, Paris)[1] thanks to a grant from the Agence nationale de la recherche (ANR-Cornafrique) between 2010 and 2012.

Medieval and early modern Ethiopian archives and electronic editing tools

Ethiopian manuscript archives is a general term encompassing administrative, juridical, and historical texts, which were produced by the Ethiopian political and religious authorities to proclaim their laws, rules and traditions. The term 'archives' is to be thought of in a very wide sense—practical writings, legal and pseudo-legal writings, local or would-be 'universal' historiography—and also as standing in juxtaposition to religious and literary texts. The producers of these documents were the royal, and, to a lesser degree, religious administrations. Private acts were issued comparatively late, from the mid-eighteenth century or a little earlier. Several thousands, perhaps even hundreds of thousands, of such documents of diverse character, constitute a coherent corpus of primary sources, so far largely under-exploited. One of the reason for this is the fact that these documents are spread in blank spaces within the liturgical and biblical manuscripts of monasteries and churches of the Ethiopian highlands, as well as in the Ethiopian manuscripts collections of the Western libraries. Ethiopian ancient archives are literally dissolved inside the libraries.[2]

Establishing ways of publishing and analysing these documents is thus part of an approach, innovative in so far as it draws on digital technologies,

1 <http://www.cn-telma.fr/>, accessed 29 May 2018.

2 See the special issue of *Northeast African Studies*, 11/2 (2011) on the topics of Ethiopian archives, edited by Anaïs Wion and gathering articles by Donald Crummey, Manfred Kropp, Claire Bosc-Tiessé and Marie-Laure Derat, Deresse Ayenachew, Anaïs Wion, and Paul Bertrand.

```
<locus target="#58va"/>
<title xml:lang="fra">
  Clause d'immunité émise par le roi Galāwdēwos (1540-59) en faveur du monastère de Wāldebbā
</title>
<note type="résumé">...</note>
<note type="scientifique">...</note>
<date type="production" notAfter="1559" notBefore="1540">1540-59</date>
<q xml:lang="gez">
  <seg ana="#invocation" type="interpretation">በአኩ፡ቴተ፡ አብ፡ ወወልድ፡ መመንፈስ፡ ቅዱስ፡</seg>
  <seg ana="#provision" type="interpretation">አነ፡ ሥራዕኩ፡ ወአው-ገጠኩ፡ ለሰብአ፡ ገጻም፡ ዘሞልድ৭</seg>
  ፡
  <seg ana="#suscription" type="interpretation">አነ፡ ንጉሥ፡ ገላውዴዎስ፡ ዘተሰመይኩ፡ አጽናፍ፡ ሰገድ፡</seg>
  <seg ana="#clauses" type="interpretation">
    ከመ፡ ኢይቅረብ፡ ፩፡ ዘኢነበረ፡ ውስተ፡ ገጻም፡ በአፈ፡ አብ፡ ወወልድ፡ መመንፈስ፡ ቅዱስ፡ ወበአፈ፡ ፲ወ፩፡ ነቢያት፡ ወ፲ወ፪፡ ሐዋርያት፡ ወበአፈ፡ አግዘእትነ፡
    ማርያም፡ ወላዲተ፡ አምላክ፡
  </seg>
  <seg ana="#clauses" type="interpretation">ከመ፡ ኢይባዕ፡ ጕልት፡ ውጉዘ፡ ይኩን፡ በሥልጣኖሙ፡ ለጴጥሮስ፡ ወጳውሎስ፡</seg>
  <seg ana="#clauses" type="interpretation">ወከመ፡ ኢይፍሐት፡ ዘንተ፡ ዘተጽሕፈ፡ ጎብ፡ ዝንቱ፡ መጽሐፈ፡ መንግሥት።።።</seg>
</q>
<q xml:lang="fr">
  À la gloire du Père, du Fils et du Saint Esprit. J'ai légiféré sous peine d'excommunication pour les gens du monastère de
  <placeName type="religious-institution" ref="place.xml#L002">Wāldebbā</placeName>
  , moi,
  <persName ref="pers.xml#p002">
    <name>Galāwdēwos</name>
    qui ait pris le nom d'
    <addName>Aṣnāf Sagad</addName>
  </persName>
  , afin que celui qui ne résiderait pas dans ce monastère n'en approche pas par la bouche du Père, du Fils et du Saint Esprit,
  et par la bouche des quinze prophètes et par la bouche des douze apôtres et par la bouche de Notre Dame Marie mère de
  Dieu. Afin qu'il ne pénètre pas dans ce
  <term>gʷelt</term>
  qu'il soit excommunié par le pouvoir de Pierre et de Paul. Et afin que l'on n'efface pas ceci qui a été écrit dans ce
  <term>maṣḥafa mangeśt</term>
  <ptr target="#note3-1"/>
  .
</q>
```

Fig. 1. XML mark-up of a document (BNFabb152#a2).

and classical by its situation within the tradition of diplomatics. The electronic publication of these documents has a number of objectives in mind: gathering, editing, translating, annotating, analysing these primary sources. EMA privileges the use of the TEI (Text Encoding Initiative) mark-up, a versatile language that can report on the different states of a text: its content, its materiality, and also the intention, notes, comments of all the authors, scribes, readers, other users, including the scholars participating to the digital edition. Indeed, one of the main scientific assets of TEI mark-up is that data can be extracted directly from the document. Encoding preserves the integrity of the text while generating layers of enriched data.[3] It offers also the possibility of interoperability, and guarantees that this publication of sources will evolve in harmony with tools and practices in international use. A starting collaboration with the project *Beta maṣāḥəft* at Hamburg has proven that interoperability is not an empty word.

Metadata architecture: sticking to the facts and promoting scientific choices

Structuring the metadata is the first scientific choice, in the limit of what the standard allows, of course. Resulting from a collaboration between myself, Anaïs Wion (IMAF-CNRS), Cyril Masset (IRHT-CNRS) and Lou Burnard (formerly consultant for TGE Adonis at CNRS, now TGIR Huma-Num), the

3 Always worth reading: Burnard 2014.

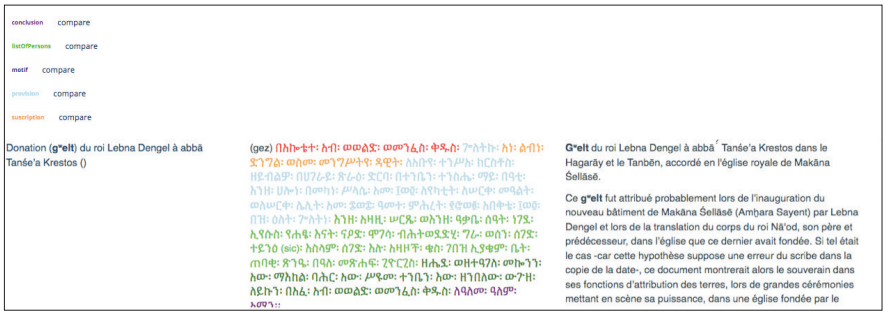


Fig. 2. View of a charter with transcription, translation, and commentary, the formulaic elements are highlighted (as appearing on <http://betamasaheft.eu> portal).

architecture chosen for the XML files in the EMA project respects the materiality of the documentation and allows the authors to create their own scientific collections. TEI deals with each level of the project, which are: text, manuscript, and corpus.

The textual document is the main item. It is the smallest semantic unit. It is encoded with a `<msItem>` and receives an `@xml:id`. Such a 'textual document' can be for instance a land charter, a list of kings, a private transfer of land property and so on. Each document is transcribed and translated. Semantic encoding (dates, place-names, names of persons, and so on) is strictly done on the text itself and not on the paratextual elements (see fig. 1, the semantic mark-up is conducted on the translation). The aim is to extract data from the primary source only, or at least to allow the readers to search through indexes and understand if the information is extracted from the primary sources or, in some cases, from elements added by the external sources. Another type of encoding wants to highlight the structure of the diplomatic discourse, using some `<seg>` tags. It concerns mainly the charters, which are very formalized documents (see fig. 1 for the XML mark-up on transcription and fig. 2 for the possible visualization). Then, the textual document is described by two different types of notes. One is a simple summary, with no other information than the one provided by the document itself (following the diplomatic classical tradition of the 'regest'). The other type, often much more elaborate, is a scientific note. This is where the analysis of the author can be displayed (fig. 3).

We decided not to impose any taxonomy for categorization (using labels such as 'charter', 'list', 'contract' and so on), but instead to mark up systematically the 'technical vocabulary' used by the documents themselves. Terms used to refer to land, to transaction, to taxation, to legal and archival practices, etc. are therefore encoded and serve as index keys to navigate the corpus (fig. 4). Here again, the philosophy is to reduce as much as possible



Fig. 3. View of the reproduction, the summary, and the scientific note accompanying a land charter (as appearing on the EMA portal).

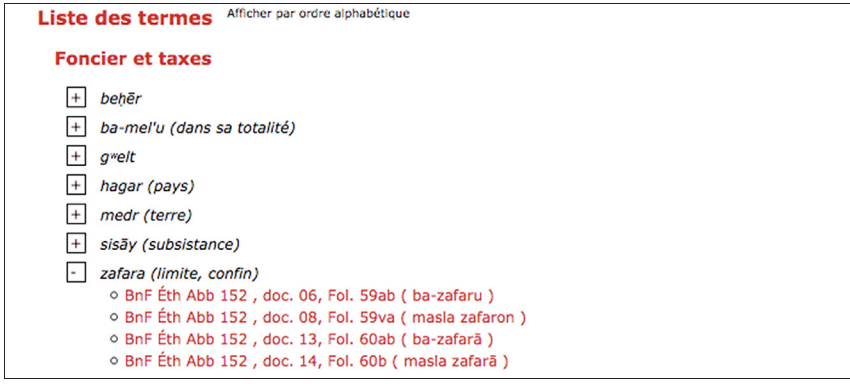


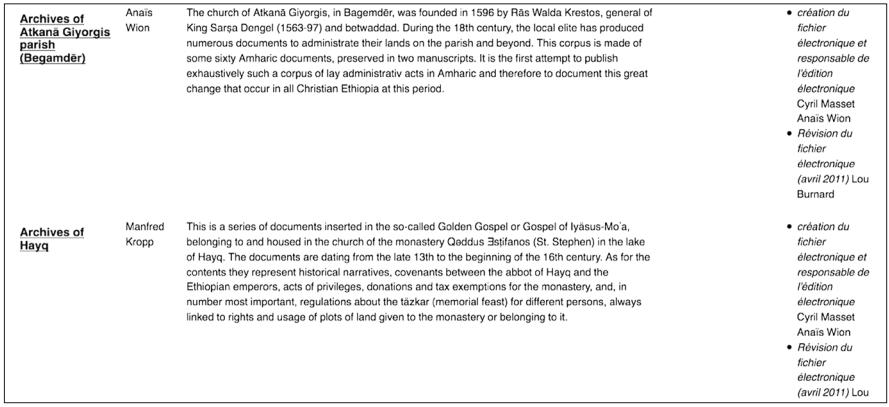
Fig. 4. The EMA filters for the 'technical vocabulary'.

external description and let the primary source display the material for their understanding.

Each of those textual items remains linked to the manuscript in which it is copied. Each manuscript is described in a separate XML file, the digital unit corresponding to the material unit. This metadata architecture was selected to stress that archival documents are spread across the manuscripts. Archival documents in Christian medieval and early modern Ethiopia have been preserved because churches and monasteries played the role of archival centres, most of the time for their own benefit but sometimes also at a regional level,

and the main media for copying and preserving pragmatic documents were pre-existing manuscripts. This feature has to be better understood, and for that it should be painstakingly documented. So far, the choices regarding the architecture of the metadata have scrupulously followed the materiality of the documentation. To put it simply: no document can be separated from the codex in which it is copied, and the codex is considered the main documentary entity.

Being an editorial project, EMA had to provide a solution going beyond the scrupulous respect of the factual reality and offer the liberty of choices and self-determination to the scholars working in it. They are considered as authors and editors of their own corpus: they gather different manuscripts with their documents, arrange them in a collection, and are in charge of their own scientific project (fig. 5).



Archives of Atkanä Giyorgis parish (Begamdër)	Anaïs Wion	The church of Atkanä Giyorgis, in Bagemdër, was founded in 1596 by Räs Walda Krestos, general of King Sarṣa Dengel (1563-97) and betwaddad. During the 18th century, the local elite has produced numerous documents to administrate their lands on the parish and beyond. This corpus is made of some sixty Amharic documents, preserved in two manuscripts. It is the first attempt to publish exhaustively such a corpus of lay administrativ acts in Amharic and therefore to document this great change that occur in all Christian Ethiopia at this period.	• *création du fichier électronique et responsable de l'édition électronique* Cyril Masset Anaïs Wion • *Révision du fichier électronique (avril 2011)* Lou Burnard
Archives of Hayq	Manfred Kropp	This is a series of documents inserted in the so-called Golden Gospel or Gospel of Iyäsus-Mo'a, belonging to and housed in the church of the monastery Qəddus Əstifanos (St. Stephen) in the lake of Hayq. The documents are dating from the late 13th to the beginning of the 16th century. As for the contents they represent historical narratives, covenants between the abbot of Hayq and the Ethiopian emperors, acts of privileges, donations and tax exemptions for the monastery, and, in number most important, regulations about the täzkar (memorial feast) for different persons, always linked to rights and usage of plots of land given to the monastery or belonging to it.	• *création du fichier électronique et responsable de l'édition électronique* Cyril Masset Anaïs Wion • *Révision du fichier électronique (avril 2011)* Lou

Fig. 5. Description of archives crediting the authorship (detail, as appearing on <http://betamasaheft.eu> portal).

TEI-XML is much more than an editing tool

Scientific expectations from EMA, which we hope will be revived by the collaboration with the *Beta maṣāḥəft* project, are numerous. The first one is to edit documents that have been used extensively for a long time, often without thinking sufficiently about what they really are, and often using a small fraction of the huge existing body of documentation. Yet, simple editing and index generation would not be a sufficient reason in order to engage in such a long and sometimes complicated process of encoding in TEI-XML. The potential this encoding offers is far too often under-used, and the wide possibilites are worth thinking about. For instance, using the tags related to the identity of persons and linking those tags with a time-line, with a space dimension, and

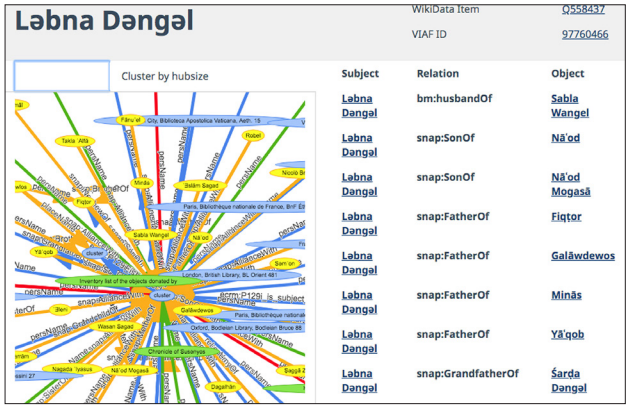


Fig. 6. Explicit and implicit relations of 'Ləbna Dəngəl', including those provided by the EMA project (as appearing on <http://betamasaheft.eu> portal).

with social networks, one could significantly improve the prosopography of Ethiopian medieval and early modern society (fig. 6). Mapping the territories that the charters and other land documents mention would also be of great help to understand Ethiopian land tenure and the political and economic issues at stake when controlling the land. But here, again, first of all one has to encode strictly what the documents say, and thus gradually gather enough data on the type of lands and tenure throughout time and space.

Generic tools for manipulating encoded texts are so far scarce, and not every scholar wants to learn XPath and XQuery in order to interrogate his or her own material. Yet there is little doubt that encoding is a rewarding intellectual activity and shall be even more so in the future. The multiplication of experiences and the dialogue between projects is for sure a sign of dynamism and it creates the conditions for building a common culture in digital humanities.

References

Burnard, L. 2014. *What is the Text Encoding Initiative? How to add intelligent markup to digital resources* (Marseille: OpenEdition Press, 2014) <http://books.openedition.org/oep/426>.

Wion, A., ed., 2011. *Northeast African Studies*, 11/2 = Special Issue: *Production, Preservation, and Use of Ethiopian Archives (Fourteenth–Eighteenth Centuries)* (2011) <http://msupress.org/journals/issue/?id=50-21D-55C>.