EAG(CENDARI): customising EAG for research purposes

Maud Medves, Laurent Romary

EAG(CENDARI): customising EAG for research purposes[1]

Maud Medves[1,2], Laurent Romary[1,2]

[1] *Inria, France*

[2] *Humboldt Universität zu Berlin, Institut für Deutsche Sprache und Linguistik, Germany*

Abstract. This paper presents the work carried out within the EU Cendari project to provide an appropriate customisation of the EAG format that would fulfil the expectations of researchers in contemporary and medieval history describing where they could find collections and documents of specific interests. After describing the general data landscape that we have to deal with in the Cendari project, we specifically address the data entry and acquisition scenario to identify how this impacts on the actual data structures to be handled. We then present how we implemented such constraints by means of a full TEI/ODD specification of EAG and point out the main changes we made, which we think could also contribute to the further evolution of the EAG setting at large. We end up providing a wider picture of what we think could be the future of archival formats (EAG, EAD, EAC-**CPF**) if we want them to be more coherent and more sustainable at the service of both archives and researchers.

Keywords: Archival standards, customisation, EAG, TEI

Historians at work

The EU CENDARI[2] is a research collaboration aiming at integrating digital archives and resources for research on medieval and modern European history. The project brings information and computer scientists together with leading historians and existing historical research infrastructures (archives, libraries and other digital projects) to improve the conditions for historical scholarship in Europe through active reflection of and considered response to the impact of the digital age on scholarly and archival practice.

In this context, we have accompanied a group of historians[3] specialised on the First World War in their information gathering activity with the objective to both identify the optimal set-up for their data entry and specify the target formats that should be used within the project. The actual content was planned to cover several levels of description of archival content:

- General information about archives that contain relevant material for WWI research;
- Specific descriptions of relevant collections within such archives, without any coverage constraint (strong sampling);

[1] This paper is also disseminated as a pre-print available from http://hal.inria.fr/hal-00959841

[2] Collaborative European Digital Archive Infrastructure (www.cendari.eu/). CENDARI is a four-year, European Commission-funded project (under the 7th Framework Programme for Research) led by Trinity College Dublin, in partnership with fourteen institutions across eight countries, to facilitate access to archives and resources in Europe for the benefit of researchers everywhere.

[3] Special thanks to Anna Bohn and Aleksandra Pawliczek from the Freie Universität Berlin team in Cendari for their very useful contributions to this work.

- Description and possibly transcription of specific items within these collections when the content may be of utmost importance for the corresponding research.

The ultimate goal in this primary data gathering activity is the production of *archival research guides*, which combine the assessment by researchers of various archival contents as to their relevance for a specified research question. Such guides are comprehensive documents that can be made available to the scholarly community at large.

A complex data space

The main challenge in Cendari is the fact that the data space it encompasses is extremely complex. Indeed we have to face a heterogeneous data world both in terms of input and of output.

Data input takes place in multiple forms: some data are entered manually, either directly using an XML editor or through editing environments such as ICA Atom. Data sources can also be provided by partner archives on the basis of so-called finding aids which in turn, depending on the digital stage level of the institution, can be available as native xml or pdf documents, but can as well only be available in print format.

Data output is also made complex because of the multiple use and publish scenarios (entries need to be readable both in html on the server and in ICA Atom environment for example). The various disseminators – The European Library is part of Cendari – or partners – APEx - also bring in their own constraints on the possible re-use of the Cendari data.

When it comes to formats Cendari has to handle (or connect to), the situation is as complex. The partners of the project ranging from archives to museums, libraries and other research projects, we are dealing with multiple data formats. The main ones are undoubtedly archival formats (EAD, EAG, EAC-CPF) as archives are the main source providers for the historians, but other have been encountered: library formats (MARC, EDM) developed by libraries or Europeana as well as formats developed by research communities (TEI, MEI). The medieval manuscripts community represents half of the historians involved in Cendari and has long been used to the TEI.

The formats mentioned above offer in turn a large standardisation spectrum and are nothing but unified, which makes the data space even more complex.

First the ubiquity of the existing standards has to be underlined. There is not one version of a standard and the various existing customisations can be compared to a large spectrum of flavours, restrictions and extensions. Each project in the field has adapted standards to fit their needs. Such customisations have so far been handle without a clear technical and editorial background for their specification.

Another crucial point in this fragmented landscape are the various levels of standardisation and maintenance: those vary from strong maintenance environments managed by solid consortiums (TEI and EAD are good examples) to a much looser standardisation strategy (it has been the case for EAG).

Finally the last challenge to tackle has to do with the transversality of some entities through the various standards. There are various levels of description, which standards correspond to: the TEI deals with document level information, EAD deals with collection descriptions and EAG with institution descriptions. The issue is to coherently encode transversal entities such as locations, people or dates, which are to be found at institution, collection and artefact or document levels. Though the information is the same (and should be extracted from a level to be integrated in another one), the granularity is very often different; elements related to an address are for example much more detailed in EAG and TEI (by means of <addressline>, <country> and other <postcode>

elements), whereas EAD and MARC have a looser way of encoding it. The difference in structuring the information may be huge, as it is the case for bibliographic fields: the three archival formats (EAG, EAD, EAC-CPF) only foresee a <descriptiveNote> element with free text, whereas both the TEI and MODS offer very structured (and deep) ways of encoding bibliographical information (see below in the EAG(Cendari) customisation part of this paper).

In the following sections, we will show the strategy we have adopted to take care of this complexity in the context of the specific data entry scenario we had in the Cendari project.

Data entry in Cendari

The Cendari data workflow has been elaborated by taking into account several constraints. There was a strong need of tracking sources and responsibility. In a project like Cendari, in which dozens of people work on the entries, it is crucial to identify who made which change, when and for which reason. Maintaining versions was also considered important, as was the need of a collaborative working environment. Finally two ways of editing were adopted: a professional XML editing environment and a user-friendlier tool to allow fine encoding for historians who felt confident with XML without excluding the less technical partners.

Following the well-known principle 'simple is beautiful', we chose a workflow requiring no heavy development, based on the three following components: oXygen[4], Subversion[5], XTF[6].

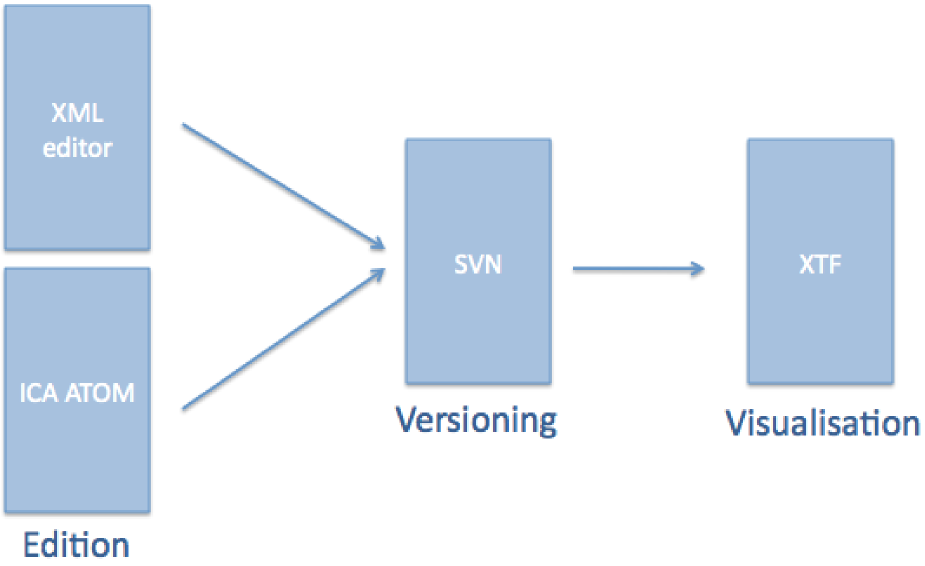


Figure 1: The Cendari data entry workflow

Once this technical workflow was agreed upon, the data entry activities (led by a group of historians) could start. The first milestone of Cendari was the elaboration of an archive directory gathering information on all institutions that were likely to provide interesting content for Cendari historians. After a phase during which historians hunted the so-called hidden archives and came back with a rich list of contacts and cooperating institutions, encoding and stocking this information was needed. EAG was an obvious candidate to address this task.

[4] The so far most advanced XML commercial editor: http://www.oxygenxml.com

[5] An Apache project faciliatation the management of versioned data: http://subversion.apache.org

[6] An open source data repository with built-in functionalities for TEI and EAD data: http://xtf.cdlib.org

The EAG model — history and scope

Encoded Archival Guide (EAG) was initially a specific initiative of the Spanish Ministry of Culture in 2002 intented to provide a format for encoding information about holders of archives. Since then, it has been largely applied in the Censo Guía de Archivos de España y Iberoamérica[7], but was never taken up by a real standardization committee. The initial proposal has been made available in the form of a Document Type Definition along with an EAG Tag Library (in Spanish).

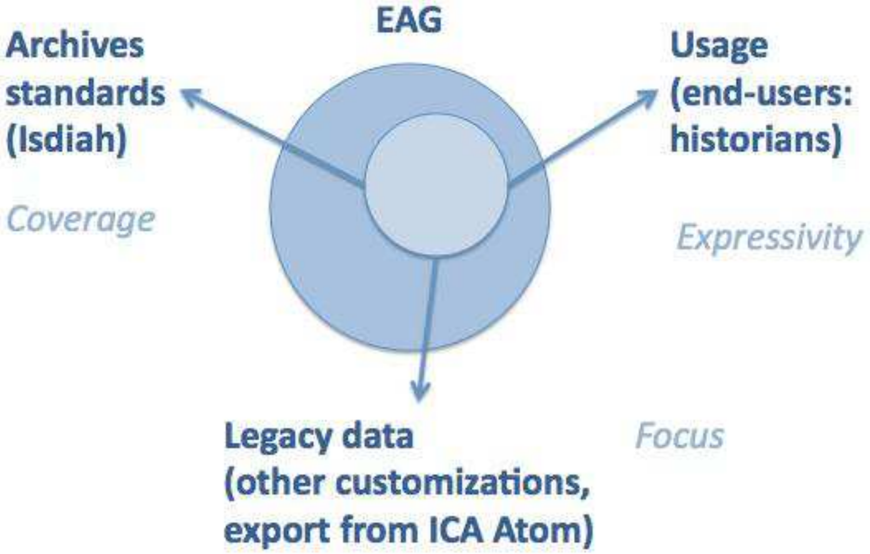
In parallel to this initiative, the International Council on Archives, through its Committee of Best Practices and Standards of the International Council of Archives (ICA/CBPS), released a more abstract standard in 2008, the International Standard for Describing Institutions with Archival Holdings (ISDIAH - http://www.wien2004.ica.org/en/node/38884), providing a precise description of all the components needed for describing holders of archives.

Interestingly, these two initiatives, which have been carried out without explicit coordination are quite aligned from a content point of view, but clearly reflect the absence of a global strategy for archival standards. Besides, the quite outdated technological background of the initial EAG proposal made it clear that we had to go a step further.

Customising EAG

In order to fulfil the requirements of the Cendari data entry workflow, and because of the somehow infancy of the EAG model, we identified the need for designing a customisation of EAG that would be a compromise between three main constraints as sketched in Figure 2.

First, we could not depart too much from the existing standards and in particular ensure our compliance with reference archival standards such as Isdiah. Second, we had to take into account the request of researchers for more expressivity in order to associate reliability information or commentaries to third-party archival information. Finally, from a pure pragmatic point of view, we had to take into account the actual legacy data we had to deal with as well as existing practices with regards to EAG in order to ensure maximal interoperability with other projects or initiatives.


EAG
Archives standards (Isdiah)
Coverage
Usage (end-users: historians)
Expressivity
Legacy data (other customizations, export from ICA Atom)
Focus

[7] http://censoarchivos.mcu.es/CensoGuia/proyecto.htm

Figure 2: constraints bearing upon the EAG(Cendari) proposal

EAG(Cendari): the customisation architecture

The work done on EAG(Cendari) relied on a comprehensive editorial platform based on the following components:

- A specification of the main EAG components implemented in the TEI ODD language;
- The TEI vocabulary proper to complement missing features in EAG;
- A feature tracker environment to record, discuss and validate the various customisation proposals made either by the technical team (i.e. the authors of this paper) or the users (the historians).

Indeed, the TEI guidelines can be seen from two different angles. First, as the basis of an XML representation format, they provide the technical constraints to control the validity of TEI conformant document instances. Second, they are delivered with an extensive prose description that informs users about the logic of the guidelines as well as the most appropriate way(s) to use them to represent specific textual phenomena[8]. Still, these two views are not split into two separated objects but indeed integrated within one single specification, from which one view and the other can be automatically generated. This mechanism, in line with the concept of literate programming (Knuth, 1992), relies in the existence of an underlying specification language named ODD (One Document Does it all), which is itself expressed in TEI.

In the TEI infrastructure, each element is thus defined as an ODD specification providing all the necessary information both to control its (XML) syntactic behaviour and to generate the corresponding documentation. Such information comprises a gloss, a definition, the technical description of its content model, the various attributes it can bear and one or several example of its usage.

In Cendari, we used this environment to facilitate the maintenance of the customisation as we made progress on it, with the advantage that we could generate on the fly and for each available version a complete set of schemas and documentation (in HTML, PDF or doc(x) formats).

In complement to this, we used an instance of Jira, a project management software, kindly provided to us by the DARIAH eInfrastructure for receiving and discussing feature requests from the historians. After an open discussion with historians, the technical team assesses if a new element or attribute should be created and the requests are implemented in an TEI/ODD document.

The context: EAG 2012

In parallel to Cendari work, the APEx project made the experience that EAG 0.2 does not fit their project purposes and started to revise EAG 0.2 to make it more compliant to ISDIAH recommendations. In order to bring together archival information from all over the continent, APEx started to revise the existing EAG created by the Spanish Ministry of Culture. In August 2012 a new EAG, revised by a consortium of 28 project partners was published by the APEx project and called EAG 2012.

Some selected new features of EAG 2012 will be presented below.

First a <location> element was introduced, wrapping information related to the physical address of an institution, in order to better structure geographical location. It allows recording different types

[8] The TEI guidelines contain in particular a wealth of examples for each element and the major constructs they allow.

of address or location per institution (visitor address vs. postal address for example). It also makes the visualisation easier through the use of geographical coordinates.

```
<repository>
    <repositoryName xml:lang="ger">Bundesarchiv Berlin-Lichterfelde</repositoryName>
    <repositoryRole>Branch</repositoryRole>
    <geogarea xml:lang="eng">Europe</geogarea>
    <location localType="visitors address" latitude="52.432423"
longitude="13.298641">
        <country xml:lang="ger">Deutschland</country>
        <firstdem xml:lang="ger">Berlin</firstdem>
        <municipalityPostalcode xml:lang="ger">12205
    Berlin</municipalityPostalcode>
        <street xml:lang="ger">Finckensteinallee 63</street>
    </location>
    <location localType="postal address">
        <country xml:lang="ger">Deutschland</country>
        <firstdem xml:lang="ger">Berlin</firstdem>
        <municipalityPostalcode xml:lang="ger">12175
    Berlin</municipalityPostalcode>
        <street xml:lang="ger">Postfach 45 05 69</street>
    </location>
[…]
</repository>
```

Secondly the introduction of a <repositories> element gives the possibility of encoding the information about several repositories (for institutions with local branches for example) within a single EAG document. In this <repositories> element, multiple <repository> children are allowed. Until this new feature, institution with head quarters and local branches (like national archives in many countries) had to record information about their branches in several EAG documents.

```
<archguide>
    <identity>
    <repositorid countrycode="DE" repositorycode="DE-0000000000001"/>
    <autform xml:lang="ger">Bundesarchiv</autform>
    </identity>
    <desc>
        <repositories>
            <repository>
                <repositoryName  xml:lang="ger">Bundesarchiv
            Koblenz</repositoryName>
                <repositoryRole>Head quarter</repositoryRole>
            […]
            </repository>
            <repository>
                <repositoryName xml:lang="ger">Bundesarchiv Berlin-
            Lichterfelde</repositoryName>
```

```
                <repositoryRole>Branch</repositoryRole>
            </repository>
        </repositories>
    </desc>
</archguide>
```

The third main feature newly introduced in EAG 2012 was the capacity to store information in several languages by making several elements repeatable. Similarly to what happened in EAD, all elements that can contain textual information allow for the attributes @lang and @script to be used. Languages and scripts used in the element can therefore be encoded. This is for example the case of <descriptiveNote> in several elements such as <repositorhist> or <holdings>. Allowing multilingualism in most of the elements is a great asset for European projects such as Cendari and Apex and favours the exchange of EAG instances at European and international levels.

Finally general contact information for the institution has been replaced by a <contact> element recording contact information for each service of an archival body. Each service can be described and contacted directly. Having entry points in most of the departments of an institution permits to identify the right person and get to him/her quicker.

```
<library question="yes">
    <contact>
        <telephone>+49 3018 7770 0</telephone>
        <email href="berlin@bundesarchiv.de">Send an e-mail</email>
    </contact>
    <webpage href="http://www.bibliothek.bundesarchiv.de/">Katalog der Bibliothek
des Bundesarchivs</webpage>
</library>
[…]
<techservices>
    <restorationlab question="yes">
    […]
        <contact>
            <telephone>+49 3018 7770 0</telephone>
            <telephone>+49 3018 7770 698</telephone>
            <email href="zwarchh@bundesarchiv.de">Send an e-mail</email>
            <email href="filmarchiv@bundesarchiv.de">Send an e-mail</email>
            </contact>
    </restorationlab>
</techservices>
```

Changes introduced in EAG(Cendari)

Similarly to APEx, Cendari identified that EAG 0.2 was missing a series of essential features for a proper description of archival institutions. We thus decided to adopt EAG 2012 when it was released and elaborated its own customisation focussing on specific needs (the focus being put on researchers). In this context, deepness was favoured over wide coverage. This induced two main consequences: a very reduced set of elements and the introduction of sourcing and referencing mechanisms.

A reduced set of elements

Compared to a standard EAG document, an EAG(Cendari) document provides as much information in the <control> and <identity> sections, but has a much limited <desc> part.

Fields relating to administrative information have been put aside to focus on fields of interest for the historians: opening hours and accessibility information have been dropped, whereas historical information and details on holdings are strongly recommended (though not mandatory).

Providing source information to EAG description

Most initially intended usages of EAG were based on the assumption that the archives, out of their internal database, would directly generate the information. In the Cendari case on the contrary, most information is gathered by researchers from existing (mostly printed) sources. Providing source information is thus essential to trace back the validity of the precise content of an EAG record, but also to identify the origin of such information, when the record is indeed a compound of several sources, as well as the researcher's own assessment of the archive.

As a matter of fact there is already an existing EAG <source> element to provide such a background, but only for the sake of qualifying the whole record. Instead of inventing a new mechanism for this purpose, we took up the existing @source attribute recently introduced within the TEI guidelines[9]. This attribute points back to a bibliographical reference or a pointer to the web site from which the information has been taken up, and indeed pointing back in our case to the <source> element[10].

The following TEI snippet illustrates this mechanism:

```
<sources>
        <source xml:id="source1">http://www.dublincity.ie</source>
        <source xml:id="source2">www.dublinheritage.ie</source>
</sources>
        [...]
<holdings xml:lang="en" source="#source1">
        <p>Dublin City Archives contains records of the civic government of Dublin from
1171 to the late 20th century. These records include City Council and committee
minutes, account books, correspondence, reports, court records, charity petitions,
title deeds, maps and plans and drawings all of which document the development of
Dublin over eight centuries.</p>
        <p>...</p>
</holdings>
```

Referencing mechanisms

Along the same requirement lines, it has been felt necessary to mark up references to Internet sources mentioned in repository descriptions. There again, the TEI guidelines offer the appropriate element (<ref>), which, by means of its @target attribute, may point to any kind of URL-defined location, as illustrated below[11]:

[9] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.source.html

[10] Note that to this end <source> has been provided with an additional @xml:id attibute.

[11] In this paper all TEI elements are marked with the prefix 'tei:' corresponding to the TEI namespace: 'xmlns="http://www.tei-c.org/ns/1.0'

```
<repositorhist>
	<p>In 1905 in preparation for the firm's 100th anniversary, a works archive is
	set up for the Krupp company (established 1811)…<p>
	<p>More information can be found on the <tei:ref
	target="http://www.thyssenkrupp.com/de/konzern/geschichte_archive_k1_2.html">Thys
	sen Krupp website</ref></p>
</repositorhist>
```

Bibliographical descriptions

One important weakness of EAG 2012 is the lack of appropriate structured bibliographical components allowing one to provide precise information about sources. We thus complemented the EAG vocabulary for Cendari with a series of bibliographical elements from the TEI vocabulary, namely <title>, <author>, <date> and <publisher>.

For instance, a simplified reference to a published article would look as follows:

```
<resourceRelation resourceRelationType="creatorOf" xml:lang="fr">
	<objectXMLWrap>
		<title xmlns="http://www.tei-c.org/ns/1.0">Le Service historique de la
	Défense, un acteur essentiel de la politique de revendication des archives mise
	en place par le ministère depuis 2009</title>
		<author xmlns="http://www.tei-c.org/ns/1.0">Michel Roucaud</author>
		<date xmlns="http://www.tei-c.org/ns/1.0">2010</date>
		<publisher xmlns="http://www.tei-c.org/ns/1.0">Revue historique des
	armées</publisher>
	</objectXMLWrap>
</resourceRelation>
```

Customisation and standardisation

As we saw in this paper, Cendari uses a profile of EAG 2012, which only slightly differs of the main schema. Both projects (APEx and Cendari) aim to establish this new EAG as a standard in the archival domain. Still, it has become clear for both projects that going further in the standardisation direction would only make sense if a better coherence between the various format oriented archival standards (EAC-CPF, EAG, EAD) could be achieved, in a context where we could not see a clear coordination in this respect, whether at technical (maintenance platform) or editorial (coherence of available features and documentation) levels.

In order to contribute to this important debate, we consider here that although EAG(Cendari) has been developed for a specific project with precise needs, it was based upon a coherent maintenance environment that could be used to develop a future standard and even to provide a comprehensive framework for the whole group of standards. The building block of such a framework should be indeed designed in such a way that a) it prevents incoherence and useless overlapping between the three, b) provides means for the community to report bugs and missing features (as possible in all open standards) and c) provides a strong technical environment to both keep track of the evolutions in the specification (versioning) and an adequate management of schemas and documentations.

As a matter of fact, even if we have focussed in this paper on the work carried out on customising EAG, the Cendari project had the opportunity to experiment something similar for EAD in order to provide collection descriptions that would fit the researchers' needs in the project. There again, we could see the advantage of using an ODD customization to both make it easy to identify the most appropriate subset for the Cendari project and complement, when necessary, the EAD vocabulary with the efficient constructs available in the TEI framework.

This experience let us devise a global vision for the future maintenance of archival standards at large, comprising EAD, EAG, but also the EAC-CPF format. The idea, depicted in Figure 3, is to have an integrated platform for the specification of all three standards based on a set of coherent editorial and technical principles:

- A joint technical committee that shares a global vision for all three standards;
- An ODD based specification for all three standards so that any shared component between the three can be maintained in a coherent way and all by-products: schemas, documentation etc. are all generated automatically from one single master specification;
- A maintenance mechanism by which requests for change in the three standards are systematically documented and discussed and periodic releases are issued;
- A general principle (inspired from the TEI guidelines) of customisation, so that projects applying archival standards can identify which subset, and possibly which extension, they are using. This should improve the comparison of actual existing flavours, with customisations being systematically documented against the reference standards;
- Editorial mechanisms for the management of feature requests, versioning, releases, that allow any user to precisely refer to the actual version of the standard he has implemented.

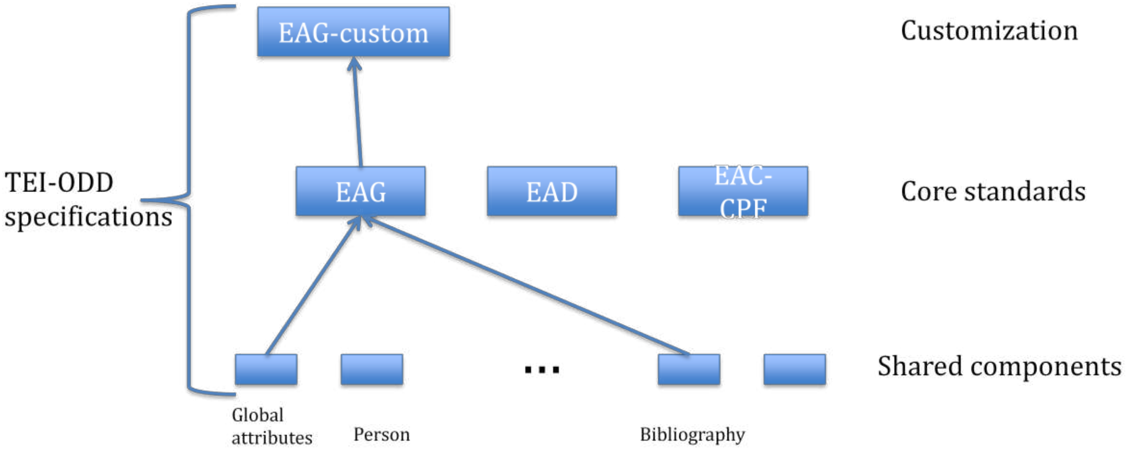


Figure 3: Maintenance architecture for archival standards

The current stage of this work is thus that we now have three ODD specifications for EAC-CPF, EAD and EAG at hand, where we systematically tried to align technical mechanisms and, when appropriate, reuse the available (and well-maintained) TEI components. Our proposal is now for the community of archivists to consider this proposal positively in order to offer better services to both archives and research.