"Multilinguality in an on-line platform for classical philology - beyond localisation of the user interface"

Cristina Vertan
University of Hamburg, Insitute for Greek and Latin Philology
cristina.vertan@uni-hamburg.de

1. Introduction

Classical philology is using increasingly repositories and adequate tools for making available to its researchers, but also to a broader public, valuable materials otherwise only available in libraries spread all over the world, and usually with restricted access. The advantages of such digital libraries are enormous, however one may consider the challenges imposed by the specificity of the domain (classical philology) to the design both of the user interface as well as to the functionality.

In comparison with digital libraries archiving modern documents, the objects in classical philology have particularities like:

- Are quite often only partially described (i.e. for one class of objects one can find one or two fields constantly mentioned for all objects.). This is mainly due to the lack of information researchers have about manuscripts
- It is almost impossible to define relations between objects which are valid for all elements inside a class (e.g. it is very often the case that not both the object description as well as object's digital form exist)
- One object contains text in several languages (Greek, Latin, at least one modern language)

Due to the above mentioned complexity up to now in classical philology we deal only with one object-type-repositories, which means that either it is a collection of manuscripts or a collection of watermarks, or collection of digitalized books, in the very best case with their descriptions (the most well known example is the Perseus digital Library [1]) We will refer to other initiatives more details in the following sections.

It is without doubt that for researcher the navigation and search possibility among various objects would be of great help in establishing interconnections, and helping decisions like e.g. establishing the date of a manuscript by means of the use of a certain watermark. Additionally such a platform is a real help for the user community, if it enables active participation of the researchers through comments, contributions, critics related to the documents on the platform. One particularity of communities in humanities, particularly in classical philology is that they have usually at least two general accepted communication languages, not only English. In classical philology, five languages (English, German, French, Spanish, Italian) are commonly accepted at conferences, journals, as well as communication means between researchers. Therefore a dedicated on-line platform should not only support all these languages (i.e. localization of the user interface) but also manage internally multilingual services. Methods from language technology and semantic web have to be used to ensure such a functionality

In this article we will present a flexible architecture that tries to embed various types of objects a classical philologist would work with, link them and offer to the users cross-lingual services.
Section 2 is giving an overview of types of data objects to be represented within the platform. Section 3 is describing the functionality of the system while section 4 refers to multilingual problems and solutions.

2. Data objects inside Teuchos platform

The Teuchos Center for Manuscript and text research [2] was set-up in 2007 at the Institute for Greek and Latin Philology at the university of Hamburg in cooperation with the Aristoteles –Archive at the Free University Berlin. It is a long-term infrastructure project, which is financed in a first phase until 2010 by the German Research Foundation in the frame of the programme „Theme-oriented information networks".

There are two main directions in which efforts of classical philologists concentrated in relation with IT. One is the construction of watermark collections, the other large repositories of digitalized manuscripts. Unfortunately up to now, to our knowledge there was no attempt to unify these collections. Our system proposes a model and encoding schema, which allows interoperability between the watermark and the manuscript collections.

Following types of documents have to be made available through the on-line research platform:

1) Manuscript descriptions: descriptions of medieval codices that have different complexity degrees from basic inventories up to deep descriptions (as for e.g. the Aristoteles Graecus ([3]). These descriptions point to parts of the corresponding manuscript but also to parts of related manuscripts.
2) Digitized versions of the manuscripts, these are image data acquired with high-resolution (sometimes also multispectral) techniques. These images have to be aligned with transcriptions (when available) and with other data.
3) Additional research data like digitized versions of watermarks graphics, biographical and bibliographical data
4) Transcriptions, and text variants for part of the manuscripts
5) Research articles, and comments created within the forum functionality

To illustrate the approach we too for the object encoding we describe below the data models for watermarks and manuscripts.

2.1. Watermarks

All medieval manuscripts referring to ancient Greek texts have watermarks. The availability of collections of watermarks is extremely important for researches as these watermarks allow often a more precise dating or establishment of the origin of the manuscript

Over the time the watermarks were catalogued on paper in various formats, but always containing a graphic (the watermark form) and a short description. On-line catalogues for medieval watermarks are already available [4]. In these on-line catalogues, the search is possible following hierarchies of watermarks motifs (graphical representation of the watermark) or following either different physical features like: size, paper structure, or characteristics of the manuscript in which they were found.

The innovative aspect of our approach is the introducing of links between different watermark instances, links that model different graphical similarities. As a consequence of the technological process for their generation, watermarks are not isolated but appear usually in pairs or even close related group of three or four. A group of several watermarks sharing the same main motif constitutes a Watermark-Motif –Object.

We are using here an Object Oriented approach, with the Watermark-Motif in the place of super-class from which each watermark instance object (unique realization of a watermark) is derived. As one motif can appear in different manuscripts we maintain a list of identical motifs. In this case the watermark was produced from the same sieve. Opposite, motifs can be very similar but produced from different sieves. In this case we will refer these motifs in a list of “similar motifs”. The realization of one motif on different pages is called instance object. In figure 1 we present the motif object while in figure 2 the motif instance object is described.

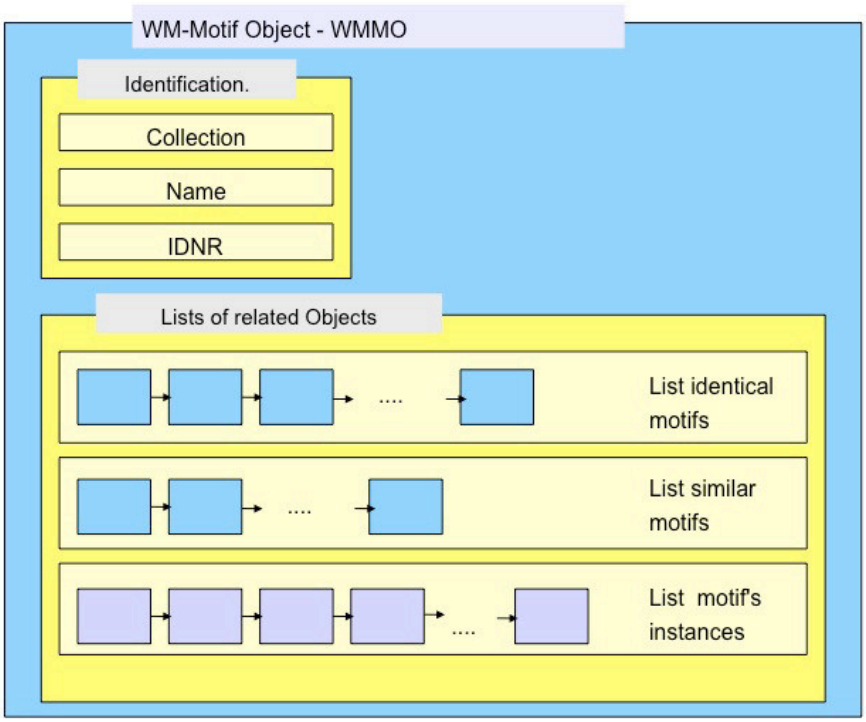

Figure 1. Watermark Motif Object

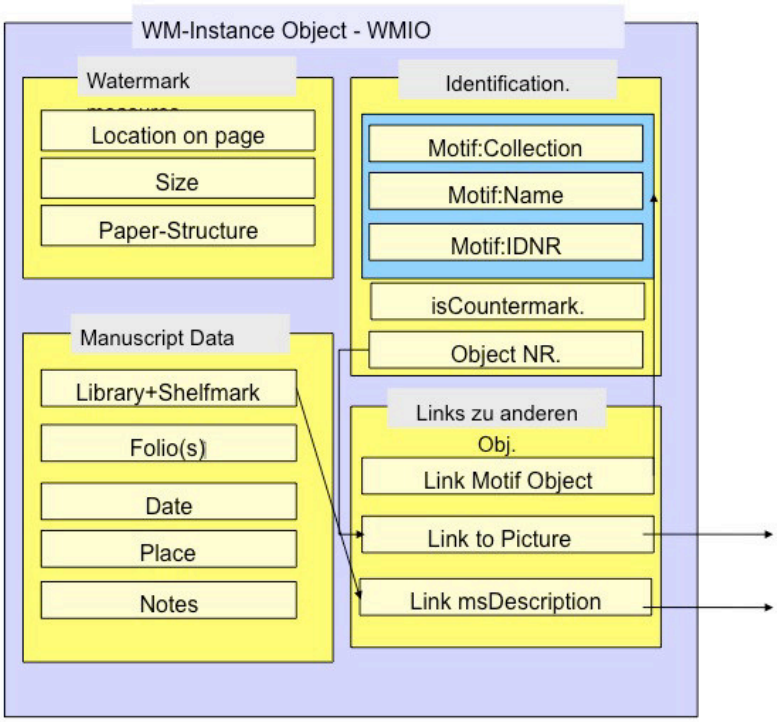

Figure 2. Watermark Motif-Instance Object

The encoding is realized in XML. We developed an XML schema, as the element “watermark” from TEI-P5 does not offer enough flexibility to record all above-mentioned information. Below we present one example of watermark instance object:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/xsl" href="wmobject.xslt"?>
  <teuwmo:teuwmObj    xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation='WatermarkObject.xsd'
   xmlns:teuwmm="http://teuchosObjects.com/watermarks/WMMotif">
  <teuwmo:wmIdent wmIsCountermark="false">

<teuwmo:wmObjId> TEU_WMDesc_Aiglem2-21.xml</teuwmo:wmObjId>
<teuwmm:wmIdentification>
        <teuwmm:wmIdnr> 21 </teuwmm:wmIdnr>
        <teuwmm:wmCollection> Harlfinger </teuwmm:wmCollection>
        <teuwmm:wmName>
               <wmNameLanguage wmLang="fr"> Aigle</wmNameLanguage>
               <wmNameLanguage wmLang="de"> Adler</wmNameLanguage>
         </teuwmm:wmName>
      </teuwmm:wmIdentification>
    </teuwmo:wmIdent>
 <teuwmo:wmManuscriptData>
    <teuwmo:msName> Vatic.1469 </teuwmo:msName>
    <teuwmo:msFolio> ff.1-72 </teuwmo:msFolio>
    <teuwmo:msDate> 1495 </teuwmo:msDate>
</teuwmo:wmManuscriptData>
<teuwmo:wmLinks>
    <teuwmo:pictureLink> Aigle-21m2.jpg </teuwmo:pictureLink>
    <teuwmo:msDescLink> Aigle-21.xml </teuwmo:msDescLink>
    <teuwmo:motifLink> Aigle.xml </teuwmo:motifLink>
 </teuwmo:wmLinks>
</teuwmo:teuwmObj>
```

As it can be observed from this example it is a common practice to record names in at least two languages for a watermark motif. The two languages are however not predefined, so any combination of two of the five languages mentioned in section 1 is possible. An ontological based approach is required in order to ensure consistency between watermark names. We will refer to this in section 3.

2.2. Manuscripts

We consider both the results of the digitalization process as well as the manuscript descriptions. We have a tolerant model in which not every manuscript description has automatically attached the collection of images for that manuscript, as well as the reverse situation where the manuscript description is missing.

A manuscript description is structured in several parts marked by keywords. These keywords (e.g. “Reklamanten”, “Kopisten”, Reklamanten”, “Inhalt”[1]) were used as identifiers for different sections of the manuscript description in the annotation process. The descriptions were encoded following a modified version of the TEI-P5 Manuscript Description Module[5]. This module was extended in order to serve to our purposes.

For example the “watermark” element as defined in TEI-P5 was inadequate for annotating the watermarks mentions in our documents. References to watermarks in manuscript descriptions are more than simple mentions of the watermark motif but include details to the place of such watermark, remarks to the respective motif etc. The annotation process was done for the moment for 100 manuscript descriptions from “Aristotle Graecus” completely automatic. Additionally we annotated automatically all remarks referring folio numbers as well as all Watermark motifs already present in our watermark collection. In this way we realized the connection between the manuscript description collection and the watermark collection. This is to our knowledge the first attempt to link different types of on-line description collections of classical philology.

[1] in our case we work with German texts. However the automatic annotation process can be easily adapted to other languages or keywords.

The annotation of folio references in the descriptions allows us as well to link digitalized versions of these folios, as soon as they become available.

3. Teuchos functionality

The system we describe intends to offer to the classical philologist a powerful tool to editing and searching and publishing materials. In this section we will describe the user scenarios we have in mind, and the system architecture that ensures the realization of these scenarios. The system architecture we refer is presented in figure 3.

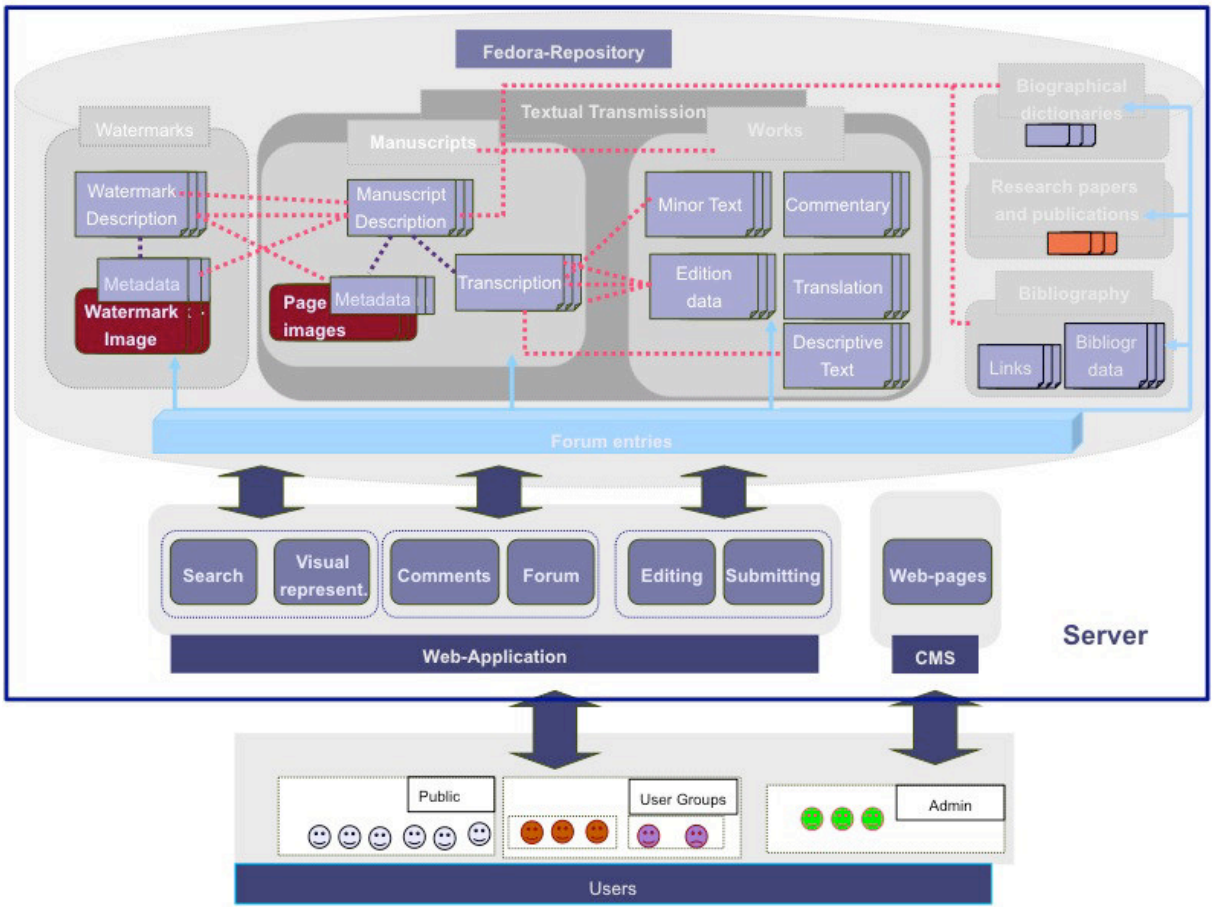

Figure 3. System Architecture

All our objects are stored on a Fedora-repository[2]- the choice of Fedora is grounded not only by the fact that it is an open source software but mainly because the stored objects can be linked following the RDF-model, i.e. gives the possibility to define semantic relations between objects.

The user interacts with the repository via a Web application that manages the editing, searching, and uploading processes. We have 3 categories of users:

- The system administrators who have full access to al parts of the system and control the content made available by the other users
- Registered users that have access to view all materials on the system can upload materials, write forum entries and upon their interest edit different material types. We envisage a hierarchy of such users having different access types to parts of the system.
- Normal users, who can only view materials declared as public.

[2] http://www.fedora.info/

The digital objects stored on the Fedora-Repository built six groups. Technically these groups are collections of Fedora-Objects.

- *Watermarks:* here are stored the watermark descriptions as well as the watermark image. Watermark images are digital graphical representations of the watermarks as they were collected in paper catalogues. Each watermark image has associated metadata in XML format, containing Dublin Core[3]-like information about the data. Each watermark description is linked to these metadata.
- The "*Textual Transmission*" group is divided in two subgroups: the manuscripts and the works
- Through "*manuscript*" we refer to one manuscript description to which we can have associated a transcription and/or a list of page images with their metadata.
- Through "*Work"* we refer to all other materials referring to one manuscript that appeared over time like: translation, different edition data, commentaries, texts about that particular manuscript. We link all these objects with the translation object.
- Two other groups the "*Biographical dictionaries*" and the "Bibliography" refer to collections of persons or woks cited in one or other manuscript description.
- A special group is dedicated *to research papers* that can refer different manuscripts.

Separating objects in such groups gives us the possibility to model two relation types:

Intra-group relations, labeled with "correspondTo", and various inter-group relations "isCitedIn", "isreferedBy", "usedFor". These relations are described as RDF-triples.

With help of these relations we are enabling a search functionality across groups, i.e. the user can for example search *"which manuscript in Collection X used the watermark Y".* To our knowledge this is the first approach in this sense, at least for classical philology.

Apart from these static objects we are implementing a forum that will give the possibility to registered users to comment any of the documents available on the repository.

4. Multilingual Aspects inside of Teuchos platform

As we mentioned in section 1 users of the Teuchos platform are speakers (or at least understand) one of five languages used inside of he community. At a first glance the straightforward consequence is the localization of the user interface in the envisaged languages. However, we claim in the following that there are deeper multilingual aspects which have to be handled and we illustrate how language technology can help.

We define three types of multilingual phenomena, occurring in our platform:
1) "Macro-document "-Multilinguality at the level of users and the uploaded multilingual documents: Therefore the platform is required not only to support uploading of documents in all these languages but also to manage their relations to one or more manuscripts in a consistent way.
2***.) "Micro-document" - Multilinguality*** at the lever of primary data to be analysed. As we already mentioned manuscripts are accompanied by modern descriptions, critical texts, which although written in modern languages (see 1) are containing often passages from the manuscript, or Latin citations. This is a real challenge when trying to process the documents automatically.
3) ***"Terminological "- Multilinguality***, as we mentioned in section 2.1. related to watermarks. Even watermarks descriptions written in one language, may declare watermarks-motifs in a variety of languages. We have to ensure that watermarks are then classified as belonging to the correct class.

To handle these three types of multilinguality we propose an ontology based approach, integrating different ontologies related to components of the system. In each of the system main components (manuscripts, watermarks, etc.) a domain specific language independent ontology ensures the correct mapping of documents on the right concept(s). Links between components are realized between the

[3] http://dublincore.org/

nodes of the ontology and not the particulars instance objects (namely the documents). The approach is represented in figure 4.

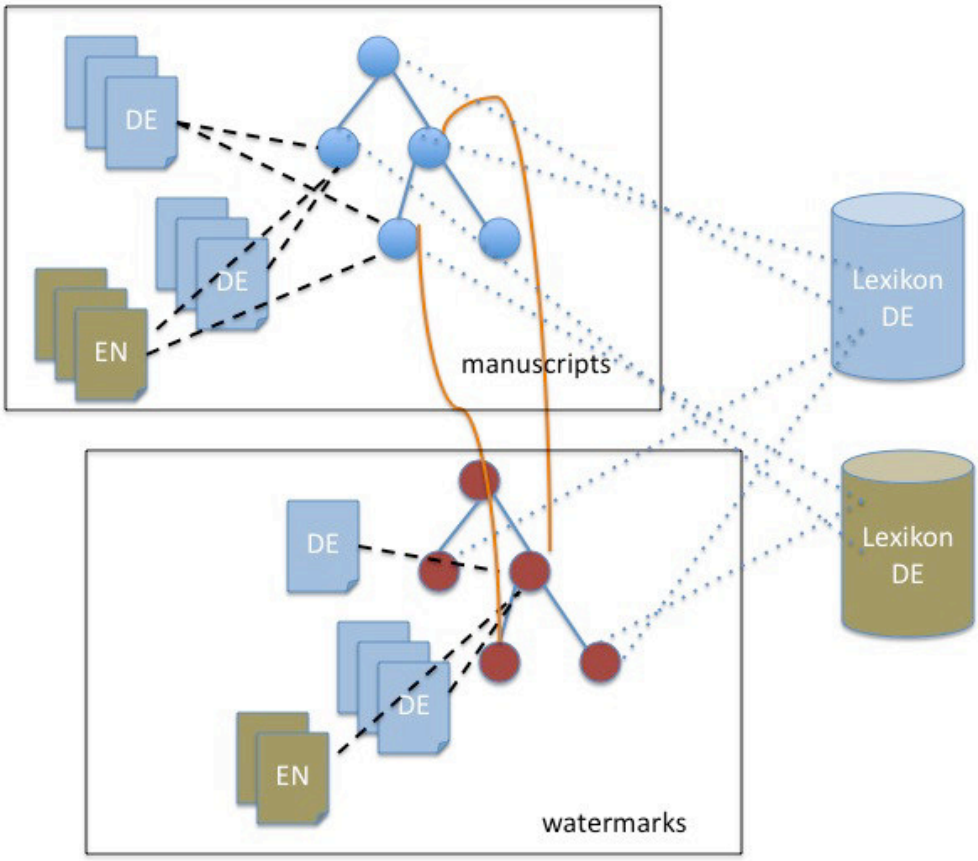

Figure 4. Ontological representation of multilingual documents.

For the mapping between the lexical materials and the ontology, the annotation of documents with concepts and the cross-lingual semantic search we intend to use the approach followed in [6].

5. Conclusions and further work

In this paper we presented an architecture for storing and accessing objects for classical philology. We introduced the main objects we manipulate, their particularities and the way are stored in our system.
We argue that the representation of multilingual objects inside the platform can be done only via an ontological approach. This will ensure both consistency for the management of various data and also enable cross-lingual retrieval.

For the moment we modeled and stored watermark- and manuscript-objects, and we interconnected these two.
We are working now on a deeper annotation of manuscript description that will allow us a more refined interconnection between objects. Through deeper annotation we understand automatic recognition and annotation of person names, indications of time (year, century etc.), titles of works.
First version of the system will be released by the end of the month. This will enable real users to post comments related to different objects and this enable us to experiment the ontological setting.

References

[1] Perseus, digital library, http://www.perseus.tufts.edu/hopper/
[2] Teuchos platform , http://www.teuchos.uni-hamburg.de/
[3] P. Moraux, Aristoteles Graecus- d. griech. Ms. d. Aristoteles, 1976, De Gruyter (Berlin, New York)
[4] on-line Watermark collections, http://www.ksbm.oeaw.ac.at/wz/wzma.php, http://watermark.kb.nl, http://www.ksbm.oeaw.ac.at/wies/
[5] Manuscript Description Model TEI-P5, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html
[6] C. Vertan, K. Simov, P. Osenova, L. Lemnitzer, A. Killing, D. Evans and P. Monachesi, Crosslingual Retrieval in an eLearning Environment, Lecture Notes in Computer Science, AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, Volume 4733/2007, Springer Berlin / Heidelberg, 2007