Report on Standardization (draft)

Laurent Romary, Piotr Banski, Jack Bowers, Emiliano Degl'innocenti, Matej Ďurčo, Roberta Giacomi, Klaus Illmayer, Adeline Joffres, Fahad Khan, Mohamed Khemakhem, et al.

▶ To cite this version:

Laurent Romary, Piotr Banski, Jack Bowers, Emiliano Degl'innocenti, Matej Ďurčo, et al.. Report on Standardization (draft). [Technical Report] Deliverable 4.2, Inria. 2017. hal-01560563

HAL Id: hal-01560563

https://hal.inria.fr/hal-01560563

Submitted on 11 Jul 2017

PARTHENOS

Pooling Activities, Resources and Tools
for Heritage E-research Networking,
Optimization and Synergies

D4.2 Report on Standardization

PARTNER(s) INRIA, CLARIN, KNAW, CNR, CNRS, CSIC, FORTH, OEAW, SISMEL, AA

DATE 26 May 2017

PARTHENOS is a Horizon 2020 project funded by the European Commission. The views and opinions expressed in this publication are the sole responsibility of the author and do not necessarily reflect the views of the European Commission.

HORIZON 2020 - INFRADEV-4-2014/2015:
Grant Agreement No. 654119

PARTHENOS
Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

REPORT ON STANDARDIZATION

Deliverable Number	D4.2
Dissemination Level	Public
Delivery date	28 April 2017
Status	Final
Author(s)	Laurent Romary (INRIA), Piotr Banski (CLARIN) Jack Bowers (OEAW) Emiliano Degl'Innocenti (CNR-OVI) Matej Ďurčo (OEAW) Roberta Giacomi (SISMEL) Klaus Illmayer (OEAW) Adeline Joffres (CNRS) Fahad Khan (CNR-ILC) Mohamed Khemakhem (INRIA) Nicolas Larrousse (CNRS) Antonis Litke (AA) Monica Monachini (CNR-ILC) Annelies van Nispen (KNAW) Maciej Ogrodniczuk (CLARIN)

Nikolaos Papadakis (AA)

Graziella Pastore (INRIA)

Stefan Pernes (INRIA)

Marie Puren (INRIA)

Charles Riondet (INRIA)

Mikel Sanz (CSIC)

Maurizio Sanesi (SISMEL)

Panayiotis Siozos (IESL-FORTH)

Reinier de Valk (DANS)

With contributions from all PARTHENOS partners

Project Acronym	PARTHENOS
Project Full title	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
Grant Agreement nr.	654119

Deliverable/Document Information

Deliverable nr./title	D4.2, Report on standardization
Document title	PARTHENOS-D4.2-Report on standardization
Author(s)	Laurent Romary (INRIA), Piotr Banski (CLARIN), Jack Bowers (OEAW), Emiliano Degl'Innocenti (CNR-OVI), Matej Ďurčo (OEAW), Roberta Giacomi (SISMEL), Klaus Illmayer (OEAW), Adeline Joffres (CNRS), Fahad Khan (CNR-ILC), Mohamed Khemakhem (INRIA), Nicolas Larrousse (CNRS), Antonis Litke (AA), Monica Monachini (CNR-ILC), Annelies van Nispen (KNAW), Maciej Ogrodniczuk (CLARIN), Nikolaos Papadakis (AA), Graziella Pastore (INRIA), Stefan Pernes (INRIA), Marie Puren (INRIA), Charles Riondet (INRIA), Mikel Sanz (CSIC), Maurizio Sanesi (SISMEL), Panayiotis Siozos (FORTH), Reinier de Valk (DANS) With contributions from all PARTHENOS partners.
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 06.02.17	First draft of chapters	Laurent Romary (INRIA) and collaborators
V 0.2 07.04.17	Revised complete draft of all chapters	Laurent Romary (INRIA) and collaborators
V 0.2a 14.04.17	Second revised draft of all chapters	Laurent Romary (INRIA) and collaborators
V 0.2b 21.04.17	Third revised draft with English corrections by Piotr Banski	Laurent Romary (INRIA) and collaborators

V 0.3 25.04.17	First version submitted to PIN	Laurent Romary (INRIA) and collaborators
V 0.4 07.05.17	Reviewed and revised for English	Sheena Bassett (PIN)
Final 12.05.17	Final version submitted to the European Commission	Laurent Romary (INRIA) and collaborators

Table of contents

Index of tables

Index of figures

1. Executive Summary

The present report reflects the second stage of the definition of the Standardisation Survival Kit (SSK) within Work Package 4 of the PARTHENOS project. On the basis of the various user scenarios presented in Deliverable 4.1, where each stage of the research process has been annotated according to the actual standards that are actually needed in order to fulfil the research task, we present here a systematic review of the activities that have to be carried out to provide support to researchers in using, but also contributing to, these standards.

2. Introduction

The present report reflects the second stage of the definition of the Standardisation Survival Kit (SSK) within Work Package 4 of the PARTHENOS project. On the basis of the various user scenarios presented in Deliverable 4.1, where each stage of the research process has been annotated according to the actual standards that are actually needed in order to fulfil the research task, we present here a systematic review of the activities that have to be carried out to provide support to researchers in using, but also contributing to, these standards.

The deliverable is organized in three sections reflecting three domains of standardisation that we see play a specific role in the research process:

- Community standards to document primary data and sources, which cover a wide variety of research community or object types standards used in managing primary input in the research process;
- Reference resources, corresponding to transversal domains used to index, categorize or organize research inputs and output;
- Protocols and procedure for the Cultural Heritage domains, which, although less related to the exchange of information proper, play an essential role in the comparison of data gathering activities in the Cultural Heritage domain.

The activities below are categorized around thematic domains that could be seen as think tanks fulfilling one or more of the standardisation stages presented in D4.1:

- Valorisation and awareness raising, when stable standards exist for fulfilling certain steps in the research process and where the emphasis should be put on providing more support to researchers in implementing them, by means of documentation, resources, examples and tools where they exist;
- Elaboration, for standards that are in a definition phase or when there is an ongoing systematic review/revision phase in the standard development;
- Preparation, for domains where lacunae have been identified, as is typically the case within task 4.4 in relation to cultural heritage analysis methods, and for which it is necessary to compile pre-standardisation documents that reflect current practices.

3. Community standards to document primary data and sources

3.1 Schema customization in TEI

3.1.1 TEI customization, an overview[1]

The Text Encoding Initiative (TEI) Guidelines are addressed at anyone who wants to interchange information stored in an electronic form. They emphasize the interchange of textual information mainly, but other forms of information such as images and sound are also addressed. The Guidelines provide a means of making explicit certain features of a text in such a way as to aid the processing of that text by computer software running on different machines (a process called markup or encoding).

The TEI Guidelines describe an encoding scheme which can be expressed using a number of different formal languages. The first editions of the Guidelines used the Standard Generalized Markup Language (SGML); since 2002, this has been replaced by the use of the Extensible Markup Language (XML). These markup languages have in common the definition of text in terms of elements and attributes, and rules governing their appearance within a text. The TEI's use of XML is ambitious in its complexity and generality, but it is fundamentally no different from that of any other XML markup scheme,

[1] Antonis Litke, Nicolaos Papadakis

and so any general-purpose XML-aware software is able to process TEI-conformant texts. The Guidelines were first published in May 1994, after six years of development involving many hundreds of scholars from different academic disciplines worldwide. During the years that followed, the Guidelines became increasingly influential in the development of the digital library, in the language industries, and even in the development of the World Wide Web itself. Since 2001, the TEI has been a community initiative supported by an international membership consortium. It was originally an international research project sponsored by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing and others.

Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. Customization is a central aspect of TEI usage and the Guidelines are designed with customization in mind. The TEI provides a number of basic, general-purpose customizations. One of the best-known of these is TEI Lite, which was originally designed as a demonstration of the customization mechanism, but has become popular as a simple TEI schema for basic encoding. Other basic customizations are listed below.

Customizations provided by the TEI Consortium :

- Lite: TEI Lite, the most widely used TEI customization; includes basic elements for simple documents.
- TEI Tite: A constrained customization designed for use by keyboarding vendors.
- TEI simplePrint: An entry-level customization, focused primarily on the needs of those encoding Western European early modern printed material.
- jTEI Article: A highly-constrained customization intended for creating journal articles, in particular for submission to the Journal of the Text Encoding Initiative
- Bare: TEI Absolutely Bare, a very barebones schema with the absolute minimum of elements.
- All: TEI with all modules included.
- Corpus: TEI for Linguistic Corpora, includes the modules for encoding linguistic corpora.
- MS: TEI for Manuscript Description, includes the elements for describing

manuscripts and complex physical aspects of documents.

- Drama: TEI with Drama, includes the TEI drama module.
- Speech: TEI for Speech Representation, includes the TEI module for spoken language.

Probably the most widely used and frequently referenced TEI customization is TEI Lite, a subset of some fifty elements claiming to satisfy the needs of 90% of TEI users, as evidenced by their actual practice in creating digital texts. The original TEI Lite (1996) was based largely on observations of existing and previous practice in the encoding of texts, particularly as manifest in the collections of the Oxford Text Archive and other collections of the period. It is, therefore, unsurprising that it seems to have become, if not a de facto standard, at least a common point of departure for electronic text centres and encoding projects world wide.

The TEI schema, the many TEI customizations and the associated guidelines are maintained with the TEI format, more precisely, with a subset called "One Document Does it all" (ODD). A quick glance at the XML source code for the TEI Lite ODD shows that it appears to be a typical TEI document, with <div> elements containing <head>s, <p>s and <list>s, containing much discursive prose, as well as <ptr> elements for cross-references and a few other specialised elements such as <egXML> for XML examples. The part of the ODD document where the specific extension schema is defined is enclosed within the <schemaSpec> element, whose contents are evaluated against the existing TEI infrastructure (element and attribute definitions together with their associate pieces of documentation), and where existing elements can be renamed or redefined, new elements and attributes may be created, and those that are spurious in the new customization can be deleted from the resulting schema. The ODD is operated on by a set of XSLT stylesheets (part of the free-standing TEI Stylesheets package that is also referenced by the Roma tool), in order to produce documentation and document grammars.

ODD, as the name indicates, is a description language that "includes the schema fragments, prose documentation, and reference documentation [...] in a single document", based on the principles of literate programming. Literate programming is a programming and documentation methodology whose "central tenet is that documentation is more important than source code and should be the focus of a programmer's activity"[2]. With ODD, semantic and structural consistency is ensured as we encode and document best practices in both machine and human-readable format.

3.1.2 Long-term archiving of TEI corpora: a proposition based on OAIS model[3]

3.1.2.1 Scope

A large number of digital resources coming from Research Communities, at least from the Humanities, use the TEI format. Considering the huge amount of work required to create these resources, there is a need to think about their preservation in order to make them reusable in the future. There is a great diversity within TEI community, which represents also the different types of described objects they deal with. Nevertheless, they share a common way of encoding by using the TEI Guidelines both for documentation and definition of their corpora. This can be considered a foundation for a good practice, but that is not enough for the digital archival community for which the main goal is to ensure that the resource should be readable and understandable in the future, say in more than 20 years, by someone who was not involved in the creation of this resource.

To fulfil this objective, the data archivist requires to verify both the technical coherence of the resource and its reusability, which means that documentation, taken in an expanded meaning of the term, should ensure that one need not find a “(TEI) Rosetta Stone” to decipher and understand it. Therefore, the idea is to identify some additional criteria, compared to those commonly used for scholarly research purposes, to reach the goal of long-term preservation of TEI corpora in conjunction with the CINES (the French National Digital Archive service - https://www.cines.fr), which will preserve them.

3.1.2.2 Technical overview

How do you define a TEI document?

Data archivists take the responsibility of ensuring that the digital format will be readable in the future, which means firstly that they can verify its conformance of the format when they receive the digital resource.

[2] Norman Walsh, Literate Programming in XML, 2002, http://nwalsh.com/docs/articles/xml2002/lp/paper.html, accessed on January 10th, 2017

[3] Nicolas Larrousse (CNRS/Huma-Num), Michel Jacobson (CNRS/LLL)

But what could TEI "conformance" be? TEI-encoded resources use TEI XML schema, so we decided to assume that all TEI documents intended to be archived should be checked against the last version of TEI schema. This means that we do not accept other namespaces, in the XML sense. If you want to add an image expressed in SVG, it will be considered as an "external" object, checked against SVG schema: you may use an "url" attribute to do so for instance.

Example: Insertion of an "external" element.

```
<figure>
	<graphic url="../Media/pic12.svg" width="356px" height="304px"/>
	<!--<figDesc>Kitten</figDesc -->
</figure>
```

This decision implies that your TEI document contains only tags defined in the TEI All schema: generally speaking, there will be some work to do on a regular TEI corpus to achieve this "TEI purity". This is a little bit binding for TEI producers, but provides a means to define technically "what is a TEI document" for data archivists. It also provides a way to process to a first syntactic validation of the document.

Use of ODD (One Document Does it all)?

Now, we have a general definition of a TEI document, what about a specific document? As described above, most TEI documents use only a subset of all the available elements and attributes, and have a specific use of some of them. To document such particular uses, an ODD document should be used. You may add some more precise constraints in the way you intend to use theses tags (e.g. "this tag is required"): From an ODD document, it is possible to generate a specific schema (e.g. in RelaxNG or Schematron), in order to check the corpus against it, and provide some human-readable documentation about your scientific choices: for instance, <p> tag will be used for a paragraph.

Therefore, to be "archivable" a TEI corpus should be accompanied by an ODD.

Example: ODD sample (theoretical example)

```
   <div>
     <p>This ODD documents a minimal TEI schema for use with the Queen's Christmas Broadcast Corpus using a bare minimum of tags and word-level linguistic analysis.</p>
     <div>
       <head>Basic text structure</head>
      <p>In this very simple schema, a document contains just a <gi>body</gi>, though this may be subdivided using nested <gi>div</gi> elements. Within the <gi>div</gi> elements only <gi>head</gi> and <gi>p</gi> elements are permitted.
  <specList>
           <specDesc key="div"/>
           <specDesc key="head"/>
           <specDesc key="p"/>
        </specList>
       </p>
     </div>
     <div>
       <schemaSpec ident="odd_example" start="TEI">
                    <!-- -->
                    <moduleRef key="core" include="p head"/>
                    <moduleRef key="textstructure" include"body div"/>
       </schemaSpec>
     </div>
   </div>
```

Extra Documentation

Data archivists are insatiable: beside the technical validation they also want some “environmental” documentation to provide the production context of the corpus.

By extra documentation, we can think of a general description of the scientific project, images of the original document (e.g. facsimile) and also different representations of the corpus based on the TEI documents (e.g. pdf, HTML) with their associated stylesheets.

Process

The CINES archival service is based on model OAIS (Open Archival Information System) which provides a general framework of organization (e.g. people, system etc.) in order to manage the preservation of information for the long term.

In short, OAIS define different entities (e.g. Producers, Consumers, Manager) communicating by means of Information Packages in the course of the whole process.

In the OAIS model, the SIP (submission information package) should be built by the data producer in conformance of recommendations made by CINES.

In this case, the introduction of a new format in CINES archival infrastructure, the decision making process is based on a continuous dialogue between TEI producers, data archivists and computer specialists from the CINES.

For the TEI format, the general structure for the package to be archived should contain:

- TEI files valid against TEI All schema
- Possibly all types of external files
- An ODD documenting the specific TEI usage
- General documentation as described previously

For this new format, we have developed a specific validation procedure to be integrated in the CINES architecture:

1) Syntactic validation of each TEI document against the last TEI All schema
2) Syntactic validation of ODD document against the last TEI All schema
3) Generation of specific schema based on ODD document
4) Validation of TEI documents against this specific schema
5) Check if all “external” document referenced in TEI documents are present in the package
6) Validation of all external documents
7) Validation of documents used as an extra documentation

3.1.2.3 Resources

- TEI Guidelines

  http://www.tei-c.org/Guidelines

- ODD

  http://www.tei-c.org/Guidelines/Customization/odds.xml

- ROMA as a tool to create ODD

  http://www.tei-c.org/Roma

- TEI GitHub

  https://github.com/TEIC

- OAIS

  https://en.wikipedia.org/wiki/Open_Archival_Information_System

- Poster presented during TEI conference in ROMA (2013)

http://digilab2.let.uniroma1.it/teiconf2013/program/posters/abstracts-posters#C146

- CINES (Centre Informatique National de l'Enseignement Supérieur)
    - General how to archive

      https://www.cines.fr/en/long-term-preservation/archive-at-cines/
    - File format

      https://www.cines.fr/en/long-term-preservation/expertises/file-format/

3.1.2.4 Ongoing efforts

We are in the final phase of implementing the validation process into the generic system used by the CINES.

We still have some technical issues: for instance ODD documents cannot always be validated against the last version of TEI-ALL due to lack of retro-compatibility.

We need to have exchanges with the TEI council regarding existing tools (e.g. Roma) and their future.

3.1.3 Project oriented EAD customization[4]

3.1.3.1 Scope

TEI ODD can be used to document data models external to the TEI environment. Several projects working with archival standards (in particular EAD) use it as well. PARTHENOS created and maintain an instance of the EAD specification in ODD, that can be used to create project oriented customizations.

With ODD, semantic and structural consistency is ensured as we encode and document best practices in both machine and human-readable format. ODD was created at first to give TEI users a straightforward way to customize the TEI schema according to their own practices and document this customization. But it is possible to describe a schema and the associated documentation of any XML format.

ODD can be processed to generate an actual schema (a DTD, an RelaxNG XML schema with embedded Schematron rules, a compact RelaxNG schema, or an XML Schema), and documentation in various formats (XHTML, PDF, EPUB, docx, odt). We used ODD to

[4] Charles Riondet (INRIA), Laurent Romary (INRIA)

completely encode the EAD standard, as well as the guidelines provided by the Library of Congress, and then derived a specific customization using Schematron rules, also described with ODD.

The solution we propose is based on a flexible and customizable methodology: It combines the complete description of the specifications in a machine-readable way, and customization facilities, easy to understand for the end-user. More important, this solution does not change the core EAD schema, but add more specific rules in a comprehensive and human-readable format, by combining the EAD schema (expressed in RelaxNG) with ISO Schematron rules. Schematron is an ISO/IEC Standard (ISO/IEC 19757-3:2016) that parses XML documents and makes "as-ser-tions about the pres-ence or ab-sence of pat-terns"[5]. It can be used in conjunction with a lot of grammar languages such as DTD, RelaxNG, etc.

3.1.3.2 Technical overview

The Table 1 overleaf presents an overview of the main elements described above, with an explanation of their particular use in the EAD-ODD.

[5] http://www.schematron.com/, accessed on November 2^{d}, 2016.

ODD element or attribute	**Definition (taken from the TEI guidelines)**	**Use in EAD ODD**	**Examples**
elementSpec/@ident (identifier)	supplies the identifier by which this element may be referenced		<elementSpec dent="archdesc"…>
elementSpec/@module	supplies a name for the module in which this object is to be declared	In our case, we have only one module, which is EAD.	<elementSpec module="EAD" …>
gloss	a phrase or word used to provide a gloss or definition for some other word or phrase	<gloss> contains the complete name of the element, as stated in the tag library.	<gloss>Appraisal Information</gloss>
desc (description)	a brief description of the object documented by its parent element, typically a documentation element or an entity	In the EAD ODD, the value of <desc> is the first half of the tag LIbrary description, which gives a formal definition of the element and which kind of information it must contain (see also the <remarks> element).	<desc>A <gi>physdesc</gi> subelement for information about the quantity of the materials being described or an expression of the physical space they occupy. Includes such traditional archival measurements as cubic and linear feet and meters; also includes counts of microfilm reels, photographs, or other special formats, the number of logical records in a database, or the volume of a data file in bytes.</desc>

ODD element or attribute	Definition (taken from the TEI guidelines)	Use in EAD ODD	Examples
classes/memberOf/@key	specifies all the classes of which the documented element or class is a member or subclass		<classes> <memberOf key="model.phrase.xml"/> </classes>
content (content model)	contains the text of a declaration for the schema documented	We copy the RelaxNG schema, but in the case where elements are defined as descendants of others elements (for instance, an XPATH such as : rng:define/rng:element/rng:element), we create an independent tei:elementSpec, and we put a corresponding rng:ref in the first element definition.	

ODD element or attribute	Definition (taken from the TEI guidelines)	Use in EAD ODD	Examples

attList/attDef	contains documentation for all the attributes associated with this element	In attDef, documentation elements such as <desc> are also used, as well as specification ones, in particular the <datatype> element which define which value the attribute can have.	<attList>
			  <attDef ident="mainagencycode">
			    <desc>A code compliant with ISO/DIS 15511 Information and Documentation International Standard Identifier for Libraries and Related Organizations (ISIL). </desc>
			    <datatype>
			      <rng:text/>
			    </datatype>
			    <remarks>
			      <p>Values should be supplied without the country code, which
			should be placed instead in the COUNTRYCODE attribute.</p>
			     </remarks>
			  </attDef>
			  <attDef ident="url">
			    <desc>An absolute (http://www.loc.gov/ead/ms99999.xml) or relative (ms99999.xml) Uniform Resource Locator.</desc>
			    <datatype>
			      <rng:text/>
			    </datatype>
			  </attDef>
			…
			</attList>

ODD element or attribute	Definition (taken from the TEI guidelines)	Use in EAD ODD	Examples
exemplum	groups an example demonstrating the use of an element along with optional paragraphs of commentary		<exemplum> <teix:egXML> <eadheader langencoding="iso639-2b" xmlns="urn:isbn:1-931666-22-9"> <eadid>[...]</eadid> <filedesc>[...]</filedesc> <profiledesc> <creation>[...]</creation> <langusage>Bilingual finding aid written in <language langcode="fre">French</language> and <language langcode="eng">English.</language> </langusage> </profiledesc> </eadheader> </teix:egXML> </exemplum>
remarks	contains any commentary or discussion about the usage of an element, attribute, class, or entity not otherwise documented within the containing element	In the EAD ODD, the <remarks> element value is the second part of the description of the EAD tag Library. The information given here are caveat (i.e. possible confusions between element), the evolution of the element specification since EAD 1.0 and the crosswalk with ISAD(G).	<remarks> <p>The <gi>physdesc</gi> element is comparable to ISAD(G) data element 3.1.5 and MARC field 300.</p> </remarks>

Table 1 – the main elements of EAD-ODD

For EHRI, we created another ODD to document the specific rules and constraints of the EHRI data model. In this new ODD file, called EHRI_EAD.odd, the generic EAD specification is imported and serves the baseline of specification. The additional constraints are added only to the elements that they refer to. Therefore, the EHRI_EAD.odd file only contains the <tei:elementSpec> and <tei:classSpec> that are modified. The merge of the two ODD files – the EAD generic and the EHRI specific – is made when we apply a transformation. The constraints that we need to add to EAD in order to ensure a smooth ingestion of descriptions in the database are of two types. First, some EAD elements are required for the good functioning of the database, for instance unique identifiers for all the descriptions (contained in <ead:eadid>). Second, some elements are made mandatory for more qualitative reasons: for instance, to ease the discoverability of its resources, EHRI requires that a minimal description in English is provided with each description unit. Another example is the fact that EHRI encourages the use of ISO standards for the representation of languages, scripts, dates, etc, as well as the interlinkage of entities, via the use of authority lists.

3.1.3.3 Resources

- EAD ODD
  https://github.com/PARTHENOSWP4/standardsLibrary/tree/master/archivalDescription/EAD/odd
- EHRI-EAD ODD
  https://github.com/EHRI/data-validations/tree/master/ODD-RelaxNG/EAD

3.1.4 Project oriented EAG customization[6]

3.1.4.1 Scope

The CENDARI (Collaborative European Digital Archive Infrastructure)[7] project was born to create a research infrastructure for World War I and medieval history, and is an example of digital ecosystem. The diverse information requirements of these two communities are met by a strategy which combines newly devised approaches with metadata and tools for data integration and ontology development.

[6] Emilane degl'Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

[7] http://www.cendari.eu

The project brings together several universities, research organisations, GLAMs and ITC labs across Europe, including collection holders, historians and digital library specialists. Although the initial emphasis of the project has been on World War I studies and medieval European culture, the methods and infrastructures constructed should be relevant to any contemporary research environment.

A core part of the project is the construction of a metadata architecture to link components of the highly complex data space occupied by historical resources. These links must encompass multiple levels of granularity from that of the institutions in which they are held, down to their constituent collections, and from there to individual items and parts of these items. These levels are inconsistently manifested across subject domains, adding further complexity to the design of an overall model for metadata.

Beyond standard collection-level description metadata, such as collection titles, dates, holding institutions, languages and component descriptions, the XML schemas used in the project include more elements which are focused closely on the specific needs of CENDARI users than would be found in generic standards such as EAD (see infra in this document). These include descriptions of lacunae (gaps in the collections), descriptions of impediments users may experience in accessing or utilizing the collection contents (such as damage to parts of it) and indications of future custodial plans for the collection.

Most CENDARI holdings schemas are mapped to EAD (Encoded Archival Description), the core standard for collection-level descriptions, but some components are used to generate the EAG (Encoded Archival Guide) records which lie above EAD in the overall hierarchy.

EAG (CENDARI flavour) is a version of EAG designed to meet the needs of CENDARI regarding Archival Guides.

3.1.4.2 Technical overview

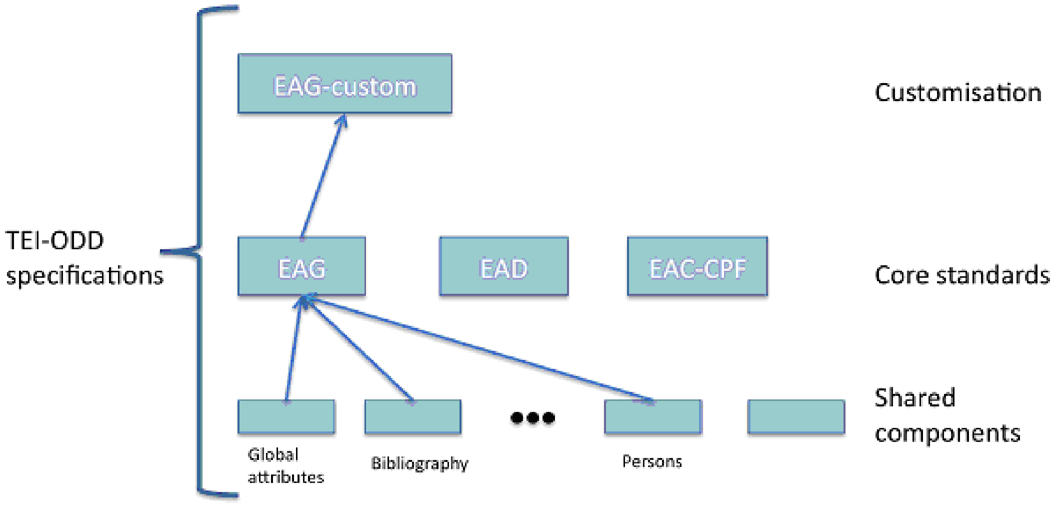


Figure 1 – The CENDARI Collection Schema (CCS)

CENDARI Collection Description

The CENDARI Collection Schema (CCS) was developed to encode detailed descriptions for collections housed by the associated cultural heritage institutions. Within the CENDARI metadata strategy collection is conceptualized as being positioned between the institution and the item. In most cases each collection will be associated with one institution that is responsible for the collection, and each collection record may also be associated with any number of item records providing detailed descriptions of items within the collection.

CCS was designed to better meet the requirements of CENDARI users than existing standards by:

- extending the standard collection-level description metadata that would be found in encodings such as EAD;
- overcoming the semantic limitations of highly descriptive elements;

The schema is written in XML (eXtensible Markup Language), a widely-used standard for metadata encoding and interchange. It aims to provide a structure to allow the most important components of collection information to be collocated and linked up as necessary. The schema defines 16 top-level components and a mechanism for linking these together using XML identifiers: in addition, every component may be identified by an

Universal Resource Identifier (URI) by which it may be linked to external resources (such as the controlled vocabularies and ontologies). Many of the elements, sub-elements, and attributes in the schema, whilst not mandatory, are nonetheless recommended for use when creating collection level records for CENDARI.

CCS has been developed for two research domains, First World War studies and Medieval History which have different requirements in terms of granularity: the collection level is of primary importance to the World War 1 community of scholars, whereas for the medievalists, the item level is the primary focus of both research and archival documentation. The extensiveness of metadata records should reflect the different user requirements, and it is expected that the collection metadata records aimed at the World War 1 community of scholars will, in most instances, be more extensive than those aimed at the medievalists. Nevertheless, some medieval collection records may require more extensive metadata than some World War 1 collection records, although even extensive metadata records will not necessarily make use of the full potential of CSS. As such, the guidelines refer to CSS Basic and CSS Full records as appropriate. A CSS Basic is a minimal collection level record that is sufficient for the identification of collections that are relevant to their research. A CSS Full is a collection level record that makes to the full use of CSS.

The 16 top-level components of a collection-level record are shown in Table 2:

Component Name	**Definition**	**Example**
1 Identifier for the collection description <collectionDescIdentifiers>	An identifier for the collection-level description itself, using any recognised format (eg. URI)	http://cendari.eu/id/collection-description/cendari-sample-1-master
2 Title for the collection-level description <collectionDescTitles>	A title for the collection-level description itself	Cendari Sample Collection 1 - Master Record
3 Holding institutions <holdingInstitutions>	Details of the archive or other organisation which hosts or administers the collection	European Imaginary Archive

Component Name	**Definition**	**Example**
4 Date <dates>	Any date associated with the archive	1922-01-01
5 Lacunae <lacunae>	Details of any material missing from the archive	Years 1923-25 are missing as a result of being eaten by mice
6 Subject coverage <subjectCoverage>	Subject terms or a prose description of the subject coverage of a collection	Middle Ages This collection is mainly centred on materials covering..
7 Languages <languages>	The languages present in the collections held by the archive	German
8 Rights Information <rightsInformation>	Intellectual property information relating to the collection	This collection is open to registered users of the archive
9 Geographic information <geogInformation>	Geographic terms associated with the collection	Germany
10 Source (provenance information) <sourceInformation>	Information on the provenance of the collection, including events in its history	John Smith donated the collection to the archive in 1922
11 Contents <contents>	A container for information on the collection as a whole and its components	
12 Relations <relations>	Any relationship between the collection and other entities (eg institutions)	
13 Usage impediment <usageImpediments>	Any factor inhibiting use of the collection	Approx. 75% of texts illegibility owing to mice damage
14 Collection future <collectionFuture>	Information on the likely future availability of the collection, or future plans for it	The collection will be maintained indefinitely at the European Imaginary Archive

Component Name	Definition	Example
15 Bibliography <bibliography>	A set of bibliographic references to literature related to the collection	
16 Record information <recordInformation>	Information on the metadata record itself, including details of its creation and changes made to it	

Table 2 – the top-level components of a CCS collection-level record

CENDARI Item Description

For item-level descriptions, CENDARI uses the MODS (Metadata Object Description Schema), supplemented by elements from the TEI P5 Manuscript Description Schema and a small number of additional elements created ex-novo by CENDARI. In skeletal outline, a record will take this form. Each component is described in the left hand column; examples are in the right column.

Component Description	Example
1. These are the schema declarations needed to invoke MODS and the MS Descriptions TEI schema.	<?xml version="1.0" encoding="UTF-8"?> <mods xmlns="http://www.loc.gov/mods/v3" xmlns:tei="http://www.tei-c.org/ns/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/mods.xsd http://www.tei-c.org/ns/1.0 http://www.tei-c.org/release/xml/tei/custom/schema/xsd/tei_ms.xsd">
2. A phrase that names the item. When the title information has not been transcribed from the item itself, the attribute supplied should be set to "yes". The title itself is given with the <title> element: this may have a lang attribute indicating the language of the title. The title element is repeatable.	<titleInfo supplied="yes"> <title lang="en"></title> </titleInfo>

Component Description	Example

3. The repeatable <name> element records any names associated with the item. The <role>/<roleTerm> element within <name> can be used to specify the form of the association with the item. It is recommended to use MARC relator codes for this if possible, in which case the type attribute of <roleTerm> should be set to code and the authority attribute set to marcrelation as in the example (if not, set type to text). Each component of the name is given in a separate <namePart> element, the type attribute of which should be set to one of family, given, termsOfAddress or date.	<name>
	    <role>
	        <roleTerm type="code" authority="marcrelator"></roleTerm>
	    </role>
	    <namePart></namePart>
	  </name>

4. The repeatable <subject> element contains subject terms associated with the item. Within the <subject> element use one of <topic> <geographic> <temporal> <titleInfo> <name> <genre> <hierarchicalGeographic> <cartographics> <geographicCode> <occupation> to indicate the type of subject being recorded.	<subject>
	    <topic></topic>
	  </subject>

5. The repeatable <language> element allows the recording of any languages used in the item. The language should be given as an ISO 639-2 code within the <languageTerm> element, with the type and authority attributes set as in the example.	<language> <languageTerm type= "code" authority="iso639-2>en</languageTerm>

Component Description	Example

6. The <originInfo> element records information on the origin of item, including dates and places associated with its creation. Five date elements may be used as appropriate: <dateCreated> <dateCaptured> <dateValid> <dateModified> or <dateOther>. Place names should be put in the <placeTerm> element within <place>	<originInfo>
	    <dateCreated></dateCreated>
	    <place>
	        <placeTerm></placeTerm>
	    </place>
	</originInfo>

7. <accessCondition> is used to record information on restrictions and condition on access to the item. This is a free-text element.	<accessCondition></accessCondition>

8. <relatedItem> may be used to provide information on digital surrogates of the item. The type attribute should be set to 'otherFormat' as in the example. The <internetMediaType> element records the format of the digital surrogates, and the <identifier> element may be used to provide its UR, URI or other identifier. If the identifier is a URI or URL, the type attribute should be set to 'uri'.	<relatedItem type="otherFormat">
	    <physicalDescription>
	<internetMediaType>tiff</internetMediaType>
	    </physicalDescription>
	    <identifier type="uri"></identifier>
	    </relatedItem>

9. Any note may be recorded in the repeatable <note> element. Its type attribute may be set to any meaningful value.	<note></note>

Component Description	**Example**

10. The <tei: msIdentifier> element records the city, repository and identification number (such as shelfmark) of the item in its <tei:settlement> <tei:repository> and <tei:idno> sub-elements respectively. A URL for the repository may be given in the ref attribute of <tei:repository>	<extension>
	  <tei:msDesc>
	    <tei:msIdentifier>
	      <tei:settlement></tei:settlement>
	    <tei:repository ref="http://repository-url.eu"></tei:repository>
	      <tei:idno></tei:idno>
	    </tei:msIdentifier>

11. The <tei: msItem> element may be used to record the incipit, explicit and colophon of the item in its <tei:incipit> <tei:explicit> and <tei:colophon> sub-elements respectively. The <tei:msItem> element is repeatable and so may record multiple components of the same item (in the case of composite items).	<tei:msContents>
	    <tei:msItem>
	      <tei:incipit></tei:incipit>
	     <tei:explicit></tei:explicit>
	      <tei:colophon></tei:colophon>
	    </tei:msItem>
	  </tei:msContents>

12. The form attribute of the <tei: objectDesc> element records the form of the item (e.g. codex).	<tei:physDesc> <tei:objectDesc form="codex">
13. The material attribute of the <tei: supportDesc> element records the material of the item is composed (e.g. paper, vellum).	<tei:supportDesc material="paper">
14. The number of leaves is recorded in the <tei:extent> element as shown.	<tei:extent>55 leaves

15. The dimensions of the item are recorded in the <tei:dimensions> element, using its sub-elements <tei:height>, <tei:width> and <tei:depth>.	<tei:dimensions>
	        <tei:height></tei:height>
	        <tei:width></tei:width>
	        <tei:depth></tei:depth>
	      </tei:dimensions>
	     </tei:extent>

Component Description	**Example**
16. The <tei:condition> element records information on the condition of the item: it contains repeatable <tei:p> elements for paragraphs of the description.	<tei:condition> <tei:p></tei:p> </tei:condition> </tei:supportDesc>
17. Information on the layout of the item is given in the <tei:layout> element within <tel:layoutDesc>. If arranged in columns, the number is given in the columns attribute.	<tei:layoutDesc> <tei:layout columns="2">In double columns</tei:layout> </tei:layoutDesc> </tei:objectDesc>
18. The <tei:musicNotation> element records information on musical notation used within the item: it contains repeatable <tei:p> elements for paragraphs of the description.	<tei:musicNotation> <tei:p></tei:p> </tei:musicNotation>
19. The <tei:scriptDesc> element records information on the script(s) used within the item: it contains repeatable <tei:p> elements for paragraphs of the description.	<tei:scriptDesc> <tei:p></tei:p> </tei:scriptDesc>
20. The <tei:decoDesc> element records information on decoration is used within the item: it contains repeatable <tei:p> elements for paragraphs of the description.	<tei:decoDesc> <tei:p></tei:p> </tei:decoDesc>
21. The <tei:bindingDesc> element records information on the item's binding(s) : it contains repeatable <tei:p> elements for paragraphs of the description.	<tei:bindingDesc> <tei:p></tei:p> </tei:bindingDesc> </tei:physDesc>

Component Description	Example
22. The <tei:provenance> elements contains sub-elements detailing previous owners or other persons associated with the item. It contains a <tei:listPerson> element within which are multiple <person> elements, one for each associated with it. Each <tei:person> element can have a role attribute to indicate whether they were the owner, curator etc of the item. Within the <tei:person> element are multiple <tei:event> elements which contain a when attribute used to record the date itself and multiple <tei:p> elements to record what form of provenance event took place.	<tei:history> <tei:provenance> <tei:listPerson> <tei:person role=""> <tei:person role="curator"> <tei:event when="1622"> <tei:p></tei:p> </tei:event>
23. The name of the person associated with each provenance item is given in the <tei:persName> element.	<tei:persName></tei:persName> </tei:person> </tei:listPerson> </tei:provenance> </tei:history> </tei:msDesc>

Component Description	Example

24. The <cen: lacunae> element is a container for multiple <cen: lacuna> elements which record details of any items missing from the archive. This is an element from the CENDARI collection-level description schema: see its entry in the documentation for this schema for a full description of its components.	<extension xmlns:cen="file:/home/richard/Dropbox/CENDARI/cendari-collection-desc.xsd">
	    <cen:lacunae>
	      <cen:lacuna lang="en"
	        type="missing component"
	typeURI="http://cendari.edu/id/lacunatypes/missingcomponent"
	        cause="mice"
	causeURI="http://cendari/edu/id/lacunacauses/mice"
	        coverageID="cendari-sample-1-component1">
	       <p>Years 1923-25 are missing as eaten by mice</p>
	        </cen:lacuna>
	    </cen:lacunae>

25. The <cen:bibliography> element is a container element for one or more <cen:biblItem> elements used for describing any bibliographic items associated with a collection or component. The type and typeURI attributes may be used to specify the type of bibliographic item. Each <biblItem> contains a <modCollection> (for multiple entries) or <mods> element, which contains the standard MODS elements for bibliographic entries.	<cen:bibliography>
	    <cen:biblItem
	      type="secondary literature"
	typeURI="http://cendari.edu/id/bibltype/secondaryliterature">
	        <modsCollection>
	          <mods>
	            <titleInfo>
	             <title>A guide to Cendari</title>
	            </titleInfo>
	            <originInfo>
	             <publisher>Imaginary Publishers</publisher>
	            </originInfo>
	          </mods>
	        </modsCollection>
	      </cen:biblItem>
	    </cen:bibliography>
	  </extension>
	</mods>

Table 3 - CENDARI Item level components

3.1.4.3 Resources

- EAD Schema (XSD file): http://www.loc.gov/ead/ead3.xsd
- EAD Index of elements: https://www.loc.gov/ead/tglib/appendix_d.html
- EAG Index of elements: http://apex-project.eu/images/docs/APEx_EAG_2012_table_20130527.pdf
- EAG Schema (XSD file): http://www.archivesportaleurope.net/Portal/profiles/eag_2012.xsd
- EAC-CPF: Schema http://eac.staatsbibliothek-berlin.de/schema/cpf.xsd
- EAC-CPF Diagram: http://eac.staatsbibliothek-berlin.de/Diagram/cpf.html
- EAG(CENDARI): customising EAG for research purposes, official document: https://hal.inria.fr/hal-00959841v2/document
- EAG CENDARI Customization: https://wiki.de.dariah.eu/x/xIDJ
- CENDARI Collection Descriptions: https://wiki.de.dariah.eu/x/9CHr
- CENDARI Item Descriptions: for item-level descriptions, CENDARI uses the MODS (Metadata Object Description Schema), supplemented by elements from the TEI P5 Manuscript Description Schema and a small number of additional elements created by CENDARI. An example of an item level description is available here: cendari-item.xml
- A skeletal documentation draft is available here: item-level-documentation0-1.doc

3.2 Specific encoding formats

3.2.1 CIDOC-CRM[8]

3.2.1.1 Scope

CIDOC-CRM has been designed and is maintained by the International Committee for Documentation at ICOM - the International Council of Museums - to help Cultural Heritage Organizations develop adequate documentation. Started as an effort to create a general data model for museums, it eventually shifted from the Entity Relation model, used by traditional databases - to adopt an object oriented approach and become a Conceptual Reference Model enabling information interchange and integration also beyond the museum community. After a transition period (2000), it eventually became an official ISO Standard ISO 21120:2006, revised as ISO 21127:2014.

[8] Emilano degl'Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

The reason behind CIDOC-CRM is to provide compatibility to data and information produced by different institutions using different data models, workflows, and terminologies. Rather that trying to fix this gap by providing yet another set of custom transformation rules, or by oversimplifying the complexity of original data, concentrating on a limited sub set of 'core' descriptors, the CIDOC reference model aims to overcome these limitations by providing a semantic reference point which will enable Cultural Heritage Organizations to render their information resources mutually compatible without sacrificing detail and precision.

The CIDOC-CRM is a standard for domain experts in cultural heritage and related domains, providing a common and extensible semantic framework, with definitions and a formal structure to describe the implicit and explicit concepts and relationships used in cultural heritage documentation, map and describe relevant information on cultural heritage objects, formulate requirements for information systems. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information participating in PARTHENOS.

Together with the PARTHENOS Entities Model - an application profile of CIDOC-CRM developed to manage the descriptions of the PARTHENOS Entities (digital objects available in the PARTHENOS Dataspace as well as services available for the users via the PARTHENOS VREs) - CIDOC-CRM is the format used to encode all the data produced and managed by the project. FORTH developed a specific component - already integrated with the D4Science Platform - to manage and support the mapping process from specific formats (EAD,TEI etc.) to CIDOC and vice versa.

3.2.1.2 Technical overview

The CIDOC-CRM is an ontology adopting an Object Oriented modelling technique (OO) serving as a basis for mediation of cultural heritage information, providing the semantic glue to integrate a vast number of disperse individual information sources - published by museums, libraries and archives - into a coherent and valuable global resource.

The scope of the CIDOC-CRM is to provide depth and quality for descriptive information intended for academic research purposes in the field of Cultural Heritage and related disciplines. Though CIDOC-CRM’s initial interest was in museums collections, its context of application was gradually extended to cover also sites and monuments relating to natural history, ethnography, archaeology, historic monuments, as well as collections of fine and applied arts, to allow the exchange of relevant information between museums, libraries and archives. The goal of enabling information exchange and integration between heterogeneous sources determines the constructs and level of detail of the CIDOC-CRM. It also determines its perspective, which is necessarily supra-institutional and abstracted from any specific local context[9]. CIDOC-CRM is specifically intended to cover contextual information: the historical, geographical and theoretical background in which individual items are placed and which gives them much of their significance and value[10].

To implement the CIDOC-CRM model, a number of elements coming from different data structures have been mapped and/or including:

- Dublin Core
- Art Museum Image Consortium (AMICO) (with the exception of data encoding information)
- Encoded Archival Description (EAD)
- MDA SPECTRUM
- Natural History Museum (London) John Clayton Herbarium Data Dictionary
- National Museum of Denmark GENREG
- International Federation of Library Associations and Institutions (IFLA) Functional Requirements for Bibliographic Records (FRBR)
- OPENGIS
- Association of American Museums: Nazi-era Provenance Standard
- MPEG7
- Research Libraries Group (RLG) Cultural Materials Initiative DTD
- Consortium for the Computer Interchange of Museum Information (CIMI) Z39.50 Profile
- Council for the Prevention of Art Theft Object ID (core and recommended categories)
- The International Committee for Documentation of the International Council of Museums (CIDOC) The International Core Data Standard for Archaeological and Architectural Heritage
- Core Data Index to Historic Buildings and Monuments of the Architectural Heritage

[9] http://www.cidoc-crm.org/scope

[10] *id.*

- CIDOC Normes Documentaires (Archeologie)/ Data Standards (Archaeology)
- English Heritage MIDAS - A Manual and Data Standard for Monument Inventories
- English Heritage SMR 97
- Hellenic Ministry of Culture POLEMON Data Dictionary

Furthermore, a number of domain specific models have been developed to better match the scientific need of communities outside the museum domain:

- FRBRoo: a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information[11].
- PRESSoo, an extension of FRBRoo, intended to capture and represent the underlying semantics of bibliographic information about continuing resources, and more specifically about periodicals (journals, newspapers, magazines, etc.)[12].
- CRMinf is a formal ontology intended to be used as a global schema for integrating metadata about argumentation and inference making in descriptive and empirical sciences such as biodiversity, geology, geography, archaeology, cultural heritage, conservation, research IT environments and research data libraries. Its primary purpose is facilitating the management, integration, mediation, interchange and access to data about reasoning by a description of the semantic relationships between the premises, conclusions and activities of reasoning[13].

During the years a number of extensions have been developed to cope with different scientific setups, and are still in the process of becoming approved by the CIDOC Special Interest Group:

- CRMarchaeo: to support the archaeological excavation process with all the realted entities and activities[14].
- CRMsci: a global schema for integration of metadata about scientific observation, measurements and processed data in descriptive and empirical sciences such as

[11] http://www.cidoc-crm.org/frbroo/

[12] http://www.cidoc-crm.org/pressoo/

[13] http://www.cidoc-crm.org/crminf/

[14] http://www.cidoc-crm.org/crmarchaeo/

biodiversity, geology, geography, archaeology, cultural heritage conservation and others in research IT environments and research data libraries[15].

- CRMgeo: a global schema for integrating spatiotemporal properties of temporal entities and persistent items[16].
- CRM Digital: an ontology to encode metadata about the steps and methods of production ("provenance") of digitization products and synthetic digital representations such as 2D, 3D or even animated Models created by various technologies[17].
- CRMba: an ontology to encode metadata about the documentation of archaeological buildings[18].

The RDF (Resource Description Framework) is the standard for Linked Data and provides an optimal adaptation for CRM described graphs. Every atomic information item is based on an oriented triple composed of two entities and a relation. Linked Data brings the technical formality for CIDOC-CRM data to be used in PARTHENOS.

As Description format, CIDOC-CRM uses the prefix “E” for entities (capital letter each word) and “P”(Properties) for relationship (lowercase). For Example :

Entity	**Relation**	**Entity**
E22_Man-Made_Object	P1_is_identified_by	E42_Identifier

Each property has a domain and a range that show from which entities originate and to which entities refer respectively. An object (for example with URI http://museum/id/object/123) can be identified as a particular type of CRM entity using the RDF statement ‘rdf:type’:

Entity	**Relation**	**Entity**

[15] http://www.cidoc-crm.org/crmsci/

[16] http://www.cidoc-crm.org/crmgeo/

[17] http://www.cidoc-crm.org/crmdig/

[18] http://www.cidoc-crm.org/crmba/

http://museum/id/object/123	rdf:type	E22_Man-Made_Object

The foundations of the CIDOC-CRM are the events happened in the past. In the model, the root of the event are the Temporal Entities (E2) that are the only that can be linked with designedly properties to Time Spans (E52), Items (E70) and Places (E53).

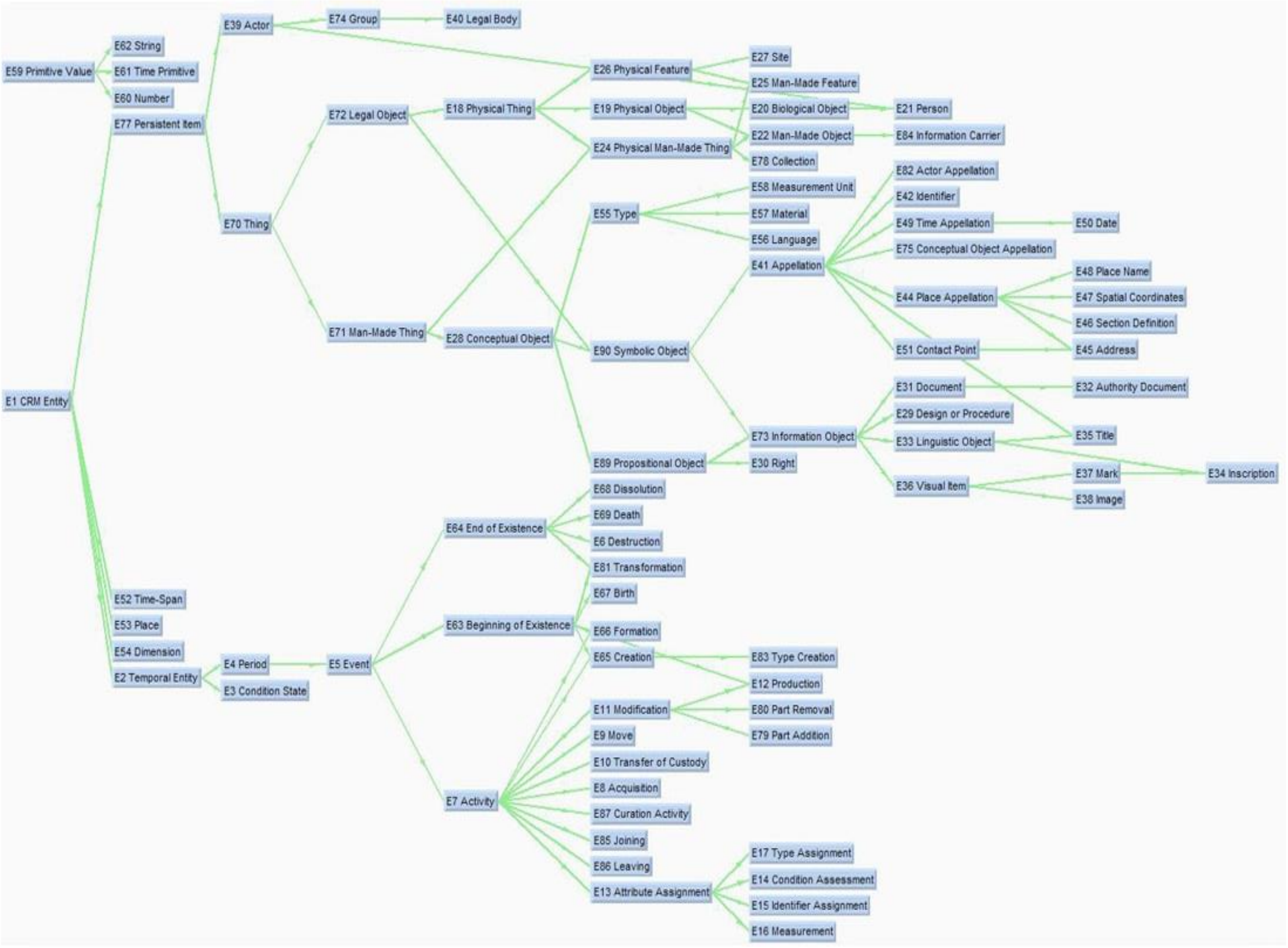
Example of actual data (in N-TRIPLE):

```
<http://authors/id/19171> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.cidoc-crm.org/cidoc-crm/E39_Actor>.

<http://authors/id/19171> <http://www.cidoc-crm.org/cidoc-crm/P1_is_identified_by> <http://names/id/7435>.

<http://names/id/7435> <http://www.w3.org/2000/01/rdf-schema#label> "Franciscus
Assisiensis".
```

Figure 2 –the CIDOC-CRM Hierarchy Class Model



3.2.1.3 Resources

- **Data repository**

http://www.cidoc-crm.org/
http://old.cidoc-crm.org/

- **Github**

https://github.com/delving/x3ml

- **Bibliography (Zotero)**

http://lists.ics.forth.gr/pipermail/crm-sig/

- **Blog entries**

http://old.cidoc-crm.org/press.htm

3.2.1.4 Ongoing efforts

3M has been developed as a web application suit to transform a structured data and other associated contextual knowledge to other schemas, in particular, the CIDOC CRM.

It assists users during the mapping definition process using a human-friendly user interface and a set of sub-components that either suggest or validate the user input.

The 3M contains several software sub-components that implement several external services. It has been developed by FORTH and is one of the PARTHENOS core tools, already integrated in the D4Science platform [19] and available for download at: http://www.ics.forth.gr/isl/3M/

All the components of the 3M have been developed as open source components in the context of the projects CultureBrokers, ARIADNE and KRHPIS – POLITEIA. More specifically, the 3M components have been released under the European Union Public Licence whereas the X3ML engine has been released under the Apache 2 licence.

Fields or elements from a source database (Source Nodes) are aligned with one or more entities described in the target schema so that the data from an entire system can be transformed. The purpose of this is typically for publication on the Web and in particular meaningful integration with other data also transformed to the same target schema.

Several participating projects (i.e.: CENDARI[20], EHRI, see infra) and databases produced

[19] https://mapping-d-parthenos.d4science.org/3M/ListEntity?type=Mapping

[20] http://www.cendari.eu

by partners (i.e.: SISMEL, etc.) are being mapped to CIDOC-CRM from their respective formats (EAD, DM2E[21], EDM[22]) using the 3M tool, to be available in the PARTHENOS content cloud.

3.2.2 Component Metadata Infrastructure (CMDI)[23]

3.2.2.1 Overview

Component Metadata Infrastructure (CMDI)[24], an ISO standard ISO-24622[25], is one of the technical pillars of CLARIN's infrastructure[26]. It features a (meta-)model to define/create and (re)use metadata schemas and at the same time a technical infrastructure to create and share these schemas as well as to create, collect and distribute actual resource descriptions (metadata records) adhering to (one of) these schemas.

Thus CMDI is specifically:

- NOT one (single) format. There is a schema[27] expressing the metamodel CMDI specification[28] and there are currently around 200 profiles[29] or schemas defined for different types of resources and different contexts.
- NOT a (single) tool. It is a set of software components forming an integrated technical infrastructure.
- The whole infrastructure is supported by a number of recommended components, guidelines and best practices, tools for validation and benchmarking, etc.

Note: Be aware of the two meanings in which the term "component" is used in CMDI:

a) the components as the core unit of the CMDI meta model

b) the software components forming the technical infrastructure.

[21] https://dm2e.eu/

[22] http://pro.europeana.eu/page/edm-documentation

[23] Matej Ďurčo (with slightly modifications by Klaus Illmayer)

[24] https://www.clarin.eu/content/component-metadata

[25] There are two parts on CMDI: ISO 25622-1:2015 describes the Component Metadata Model: https://www.iso.org/standard/37336.html, whereas ISO/AWI-24522-2 - which is currently under development - describes the component metadata specific language: https://www.iso.org/standard/64579.html

[26] https://www.clarin.eu

[27] https://infra.clarin.eu/CMDI/1.x/xsd/cmd-component.xsd

[28] https://www.clarin.eu/cmdi1.2-specification

[29] https://catalog.clarin.eu/ds/ComponentRegistry/

CMDI model

At the core of CMDI is a modular meta model allowing the definition of custom schemas. It uses concepts (data categories) defined in the CLARIN Concept Registry[30] for semantic interoperability.

Brief summary of the main concepts of the CMDI model:

- **Component** - a reusable container to describe certain aspect of a resource
  - Contains “elements” = description fields
  - Can be recursive (can contain other components)
  - E.g.: address, author, project, technical information, …
- **Profile** - a special (top-level) component for describing certain kind of resource. An XSD file is derived from it. Profiles can be “private”, which means that a) they do not appear in the UI of the Component Registry, b) they can still be changed by their owner (as opposed to public profiles, which are frozen and changes can be only done to a new version). Note, that there are a number of records in the Virtual Language Observatory (VLO)[31] based on such private profiles. This is not much of a problem as the schemas of these private profiles still can be retrieved via the REST-API of the Component Registry.
- **Element** - the actual field carrying the value
  - Should have a link to a concept (@ConceptLink) for explicit semantics
  - Can have a closed set of values (enumeration)
  - E.g.: lastName, gender, title, sizeUnit, iso-639-3-code ...
- **Concept** - independent of the structural information, in CMDI, a set of concepts for describing language resources has been formalized (see the Clarin Concept Registry). These concepts are used to annotate Components and Elements to indicate their semantics. This mechanism ensures a first level of semantic interoperability, by clustering/linking together all fields in all profiles annotated with the same concept, irrespective of their actual name or structural position.
- **Metadata record** - an instance of a profile/schema - an XML record describing a specific resource.

[30] https://openskos.meertens.knaw.nl/ccr/

[31] Please see below “VLO”.

CMDI Technical infrastructure

CLARIN infrastructure consists of a number of central services and the CLARIN Centers, the actual content (and metadata) providers.

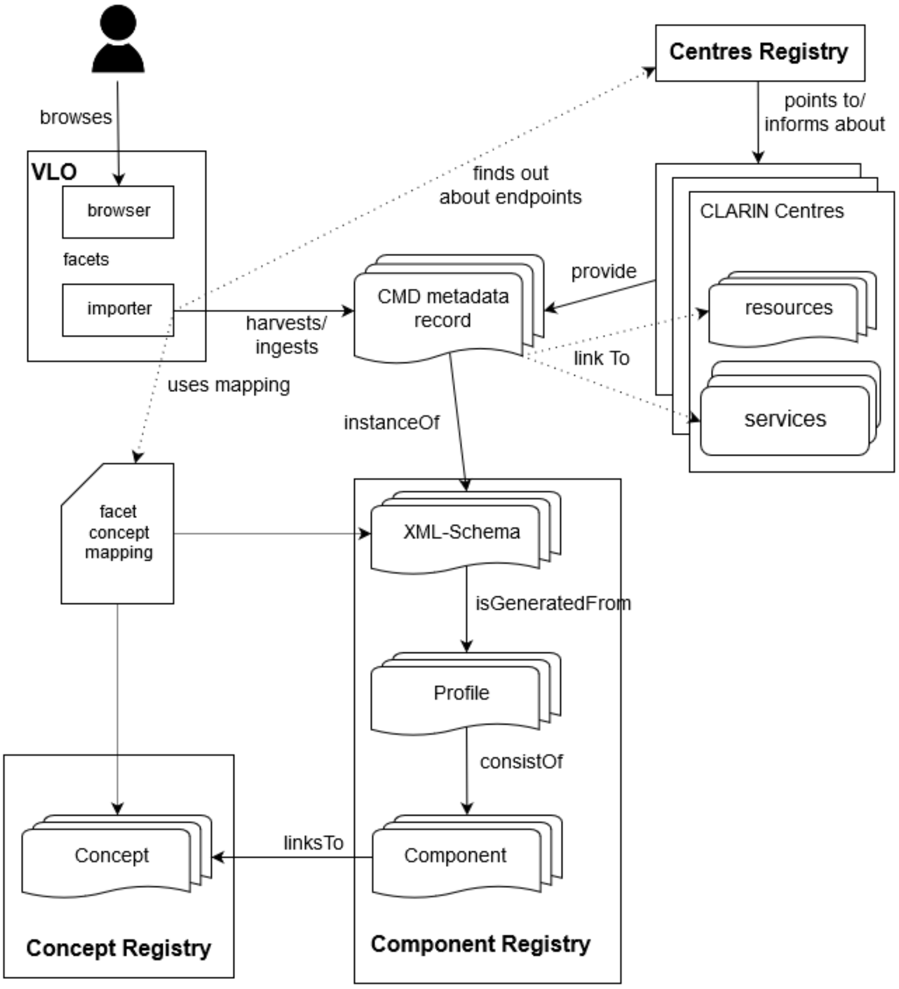


Figure 3 - the CLARIN Infrastructure

Centres Registry

CLARIN Centres Registry[32] is the primary starting point to explore the CLARIN network of centres. It offers the authoritative information about all CLARIN Centres including contact and available endpoints. In particular, information on OAI-PMH endpoint is used by the VLO[33] harvester for auto-configuration.

[32] http://centres.clarin.eu/

[33] Please see below “VLO”.

CLARIN Component Registry

The CLARIN Component Registry[34] is a registry for CMDI components and profiles. It provides a web interface for browsing, creating and publishing components and profiles and REST web service for browsing. Only authorized users can create and publish new artifacts. A metadata schema (an XSD file) is automatically generated from the definition of a profile and publicly available via the REST-API (e.g. TextCorpusProfile.xsd[35]).

CLARIN Concept Registry

The CLARIN Concept Registry (CCR)[36] is a registry of concepts relevant for the domain of language resource. These concepts form the semantic layer of CMDI (as explained above). The data model of the registry is based on SKOS. All concepts are identified by a persistent identifier (PID). The registry features an editor for curating the concepts, a faceted browser[37] for exploring them and a REST API[38] for programmatic access. It is the successor of ISOcat - Data Category Registry[39].

Virtual Language Laboratory (VLO)

The Virtual Language Observatory (VLO)[40] is a web based browser/catalog for (metadata of) CLARIN resources. It consists of the following main software components:

- a dedicated central CLARIN harvester[41] harvesting regularly (~ 1/week) all CLARIN centres via OAI-PMH protocol. The harvester is auto-configurable based on information in the *Centres Registry*.
- VLO-importer[42] transforms CMDI records into “Solr documents” *based on the facetConceptMapping*[43] (see section below) and pushes them to Solr indexer
- Apache Solr[44] in the backend as an indexing & querying engine
- a web application[45] - a faceted browser relying on the Solr API for faceting and querying the indexed data.

[34] https://catalog.clarin.eu/ds/ComponentRegistry

[35] https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1290431694580/xsd

[36] https://openskos.meertens.knaw.nl/ccr/

[37] https://www.clarin.eu/ccr

[38] https://openskos.meertens.knaw.nl/ccr/api/

[39] http://isocat.org/

[40] https://vlo.clarin.eu/, see also clarin-account on github for source code: https://github.com/clarin-eric/VLO/

[41] https://vlo.clarin.eu/data/

[42] https://github.com/clarin-eric/VLO/tree/master/vlo-importer

[43] https://github.com/clarin-eric/VLO/blob/master/vlo-commons/src/main/resources/facetConcepts.xml

[44] http://lucene.apache.org/solr/

3.2.2.2 Usage example

VLO facet mapping

The transformation process of a CMD record into a Solr document is called facet mapping. Although CMDI is a very flexible meta-format allowing for a wide variety of metadata structures, defining the mapping on structural level (identifying individual XPaths) is next to impossible. However, CMDI has the built-in concept-based semantic interoperability layer, designed exactly to allow for semantic mapping, independently of the structure. The mapping mechanism relies on the semantic annotations of the elements in the CMDI schemas with concepts defined in the CCR into VLO facets.

The following is an example snippet of a CMDI profile specification (AnnotationCorpusProfile[46] as CMDI spec):

```
<ComponentSpec isProfile="true" … >
        <Header>
                <ID>clarin.eu:cr1:p_1357720977520</ID>
                <Name>AnnotatedCorpusProfile</Name>
        </Header>
        <Component
                name="AnnotatedCorpusProfile"
                CardinalityMin="1" CardinalityMax="1">
                <Component
                        name="GeneralInfo"
                        ComponentRef="clarin.eu:cr1:c_1359626292113"
                        CardinalityMin="1"
                        CardinalityMax="1">
                        <Element
                                name="ResourceName"
                                ConceptLink="http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-
                        ebfd-241c0464e7e5"
                                ValueScheme="string"
                                CardinalityMin="1"
                                CardinalityMax="unbounded"
                                Multilingual="true"/>
                        ...
```

In the corresponding XSD[47], this (Element definition) translates to:

```
<xs:element
        name="ResourceName"
        maxOccurs="unbounded"
        minOccurs="1"
        cmd:ConceptLink="http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5">
```

[45] https://github.com/clarin-eric/VLO/tree/master/vlo-web-app

[46] https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1357720977520/xml

[47] https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1357720977520/xsd

The mapping process proceeds in two steps:

In the first step, a mapping is created on the schema (profile) level. This mapping is based on the facetConcepts.xml[48] file which contains the list of facets and a set of corresponding concepts for each of them. During the mapping process, we try to match one or more concepts for each facet in the profile's schema (checking for the **cmd:ConceptLink** in the XSD file). If some concept is matched, we say that the profile covers the facet and an *XPath to the matching element* is constructed. (Multiple XPaths are possible per single concept. Also, different facets could address the same concepts.) The product of the mapping is a set of facets and the related XPaths. This is done for every profile. (The facetConcepts.xml also contains XPaths called "fallback patterns" for some of the facets that are used in the case that none of the concepts is matched in the profile.)

In the second step, the newly created facet-to-XPath mapping is used to extract values from the CMD records. For each facet, the corresponding XPaths are evaluated against the CMD records to obtain the value (or values in case the given facet supports multiple values) and to construct the Solr document that is finally sent for indexing to Solr.

For more information about the mapping see van Uytvanck, D 2013: How does the mapping to the VLO facets work?[49].

3.2.2.3 Further reading

CLARIN ERIC, Frequently Asked Questions - Metadata in CLARIN: basics, https://www.clarin.eu/faq-page/273 (last accessed April 2017).

CMDI Task Force 2016, CMDI 1.2 specification, https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf (last accessed April 2017).

Goosen, T, Windhouwer, M, Ohren, O, Herold, A, Eckart, T, Durco, M & Schonefeld, O 2015, CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. in J Odijk (ed.), *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands.*, 116:004, Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköpings universitet, Linköping, pp. 36-53,

[48] https://raw.githubusercontent.com/clarin-eric/VLO/master/vlo-commons/src/main/resources/facetConcepts.xml

[49] https://www.clarin.eu/faq/how-does-mapping-vlo-facets-work

https://pure.knaw.nl/portal/en/publications/cmdi-12-improvements-in-the-clarin-component-metadata-infrastructure%2891536b93-31cb-4f4a-8125-56f4fe0a1881%29.html (last accessed April 2017).

Wittenburg, P, van Uytvanck, D 2012: The Component Metadata Initiative (CMDI), in: CLARIN-D AP 5, *CLARIN-D User Guide*, https://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml (last accessed April 2017).

3.2.3 Alto-XML[50]

3.2.3.1 Scope

ALTO (Analyzed Layout and Text Object) is an open XML Schema developed by the EU-funded METAe project group for use with the Library of Congress' Metadata Encoding and Transmission Schema (METS). However, ALTO instances can also exist as a standalone document used independently of METS.

The standard was initially developed for the description of text OCR and layout information of pages for digitized material. The goal was to describe the layout and text in a form to be able to reconstruct the original appearance based on the digitized information - similar to the approach of a lossless image saving operation.

ALTO stores layout information and OCR recognized text of pages of any kind of printed documents like books, journals and newspapers. METS provides metadata and structural information while ALTO contains content and physical information.

CCS Content Conversion Specialists GmbH maintained the ALTO standard until 2009. This company was involved with ALTO during the METAe project. From 2009, the Library of Congress (LC) Network Development and MARC Standards Office became the official maintenance agency for the ALTO XML Schema. At that time LC set up an Editorial Board to help shape and advocate for ALTO. The Board thus oversees maintenance of the ALTO XML Schema and helps foster usage in the digital library community.
ALTO XML is also being used by Archival institutions to enhance access and fulltext findability of digitized Archives.

[50] Annelies vans Nispen (KNAW-NIOD)

When large digitized corpora with ALTO-XML become full-text searchable, Historical Research, Language studies profit. It eg. makes Named Entity Recognition and Text-mining possible.

3.2.3.2 Technical overview

ALTO is a standardized XML format to store layout and content information. Each ALTO file contains a style section where different styles (for paragraphs and fonts) are listed.

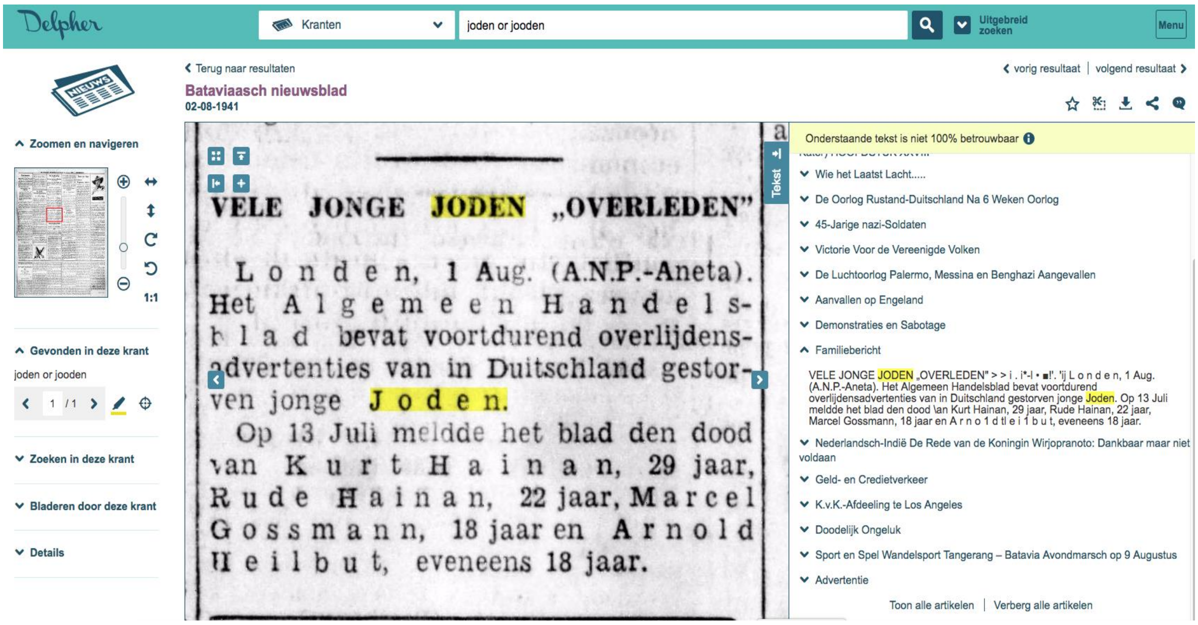
The Alto schemas are maintained and updated by the Library of Congress, all information is available at https://www.loc.gov/standards/alto/. The current version is Schema Version 3.1 [51]

ALTO schemas are updated by whole numbers upon making changes that break backward compatibility (version 1 to version 2), and decimals for changes that will not (2.0 to 2.1).
The namespace itself will also only change on major versions (ns-v2 to ns-v3).
The file location for the schemas will follow this pattern: Each major version will have its own subdirectory at www.loc.gov/alto, and the current schema (minor version) will be called alto.xsd in that directory.

ALTO-XML is used in large digitization efforts such as the Digitization of Dutch Newspapers available through Delpher and Europeana Newspapers. In the Netherlands, the use of ALTO-XML is becoming more widespread within archives. Recent digitized archives use ALTO-XML to enhance accessibility and fulltext findability of archives.

Two examples from the Netherlands:

[51] https://www.loc.gov/standards/alto/news.html#3-1-released



Bataviaasch nieuwsblad
02-08-1941

VELE JONGE JODEN „OVERLEDEN"

Londen, 1 Aug. (A.N.P.-Aneta). Het Algemeen Handelsblad bevat voortdurend overlijdensadvertenties van in Duitschland gestorven jonge Joden.

Op 13 Juli meldde het blad den dood van Kurt Hainan, 29 jaar, Rude Hainan, 22 jaar, Marcel Gossmann, 18 jaar en Arnold Heilbut, eveneens 18 jaar.

Onderstaande tekst is niet 100% betrouwbaar

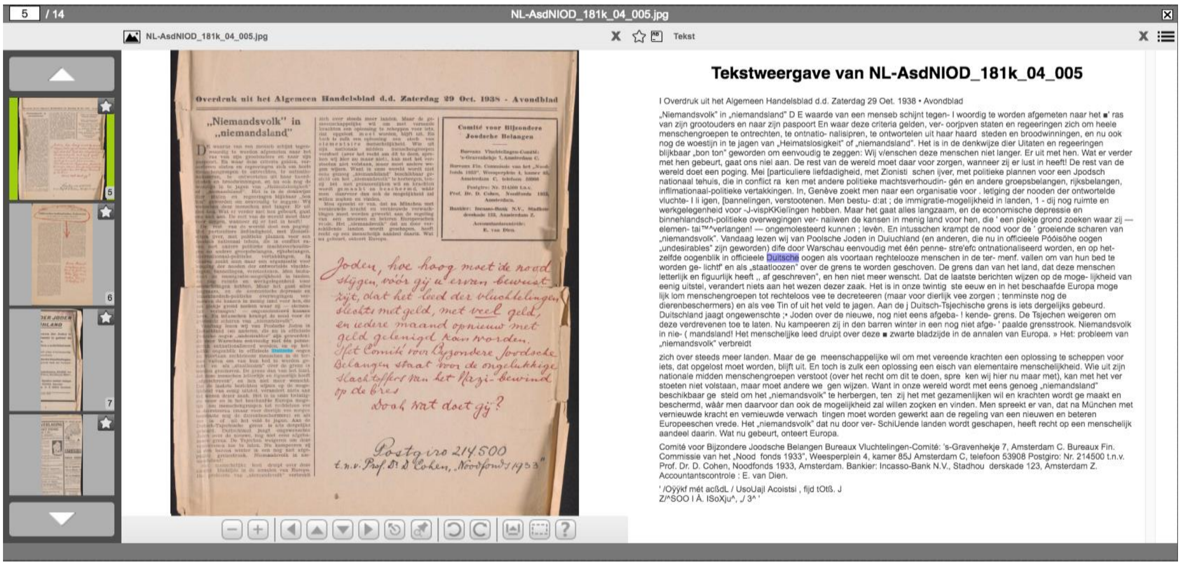
- Wie het Laatst Lacht.....
- De Oorlog Rustand-Duitschland Na 6 Weken Oorlog
- 45-Jarige nazi-Soldaten
- Victorie Voor de Vereenigde Volken
- De Luchtoorlog Palermo, Messina en Benghazi Aangevallen
- Aanvallen op Engeland
- Demonstraties en Sabotage
- Familiebericht

VELE JONGE JODEN „OVERLEDEN" > > i . i*-l • ■!'. 'ij L o n d e n, 1 Aug. (A.N.P.-Aneta). Het Algemeen Handelsblad bevat voortdurend overlijdensadvertenties van in Duitschland gestorven jonge Joden. Op 13 Juli meldde het blad den dood \an Kurt Hainan, 29 jaar, Rude Hainan, 22 jaar, Marcel Gossmann, 18 jaar en A r n o 1 d tl e i 1 b u t, eveneens 18 jaar.

- Nederlandsch-Indië De Rede van de Koningin Wirjopranoto: Dankbaar maar niet voldaan
- Geld- en Credietverkeer
- K.v.K.-Afdeeling te Los Angeles
- Doodelijk Ongeluk
- Sport en Spel Wandelsport Tangerang – Batavia Avondmarsch op 9 Augustus
- Advertentie

Toon alle artikelen | Verberg alle artikelen

Figure 4 – Example 1 of ALTO-XML in use: "Familiebericht". "Bataviaasch nieuwsblad". Batavia, 02-08-1941. Geraadpleegd op Delpher op 30-03-2017, http://resolver.kb.nl/resolve?urn=ddd:011221931:mpeg21:a0060



Overdruk uit het Algemeen Handelsblad d.d. Zaterdag 29 Oct. 1938 - Avondblad

„Niemandsvolk" in „niemandsland"

Joden, hoe hoog moet de nood stijgen, vóór gij u ervan bewust zijt, dat het leed der vluchtelingen, slechts met geld, met veel geld, en iedere maand opnieuw met geld gelenigd kan worden. Het Comité voor Bijzondere Joodsche Belangen staat voor de ongelukkige slachtoffers van het Nazi-bewind op de bres. Doch wat doet gij?

Postgiro 214500 t.n.v. Prof. Dr. D. Cohen, Noodfonds 1933

Tekstweergave van NL-AsdNIOD_181k_04_005

I Overdruk uit het Algemeen Handelsblad d.d. Zaterdag 29 Oet. 1938 • Avondblad

„Niemandsvolk" in „niemandsland" D E waarde van een menseb schijnt tegen- I woordig te worden afgemeten naar het ■' ras van zijn grootouders en naar zijn paspoort En waar deze criteria gelden, ver- oorjpven staten en regeeringen zich om heele menschengroepen te ontrechten, te ontnatio- nalisipren, te ontwortelen uit haar haard steden en broodwinningen, en nu ook nog de woestijn in te jagen van „Heimatslosigkeit" of „niemandsland". Het is in de denkwijze dier Uitaten en regeeringen blijkbaar „bon ton" geworden om eenvoudig te zeggen: Wij v/enschen deze menschen niet langer. Er uit met hen. Wat er verder met hen gebeurt, gaat ons niel aan. De rest van de wereld moet daar voor zorgen, wanneer zij er lust in heeft! De rest van de wereld doet een poging. Mei [particuliere liefdadigheid, met Zionisti schen ijver, met politieke plannen voor een Jpodsch nationaal tehuis, die in conflict ra ken met andere politieke machtsverhoudin- gén en andere groepsbelangen, rijksbelangen, inflimationaal-politieke vertakkingen. In, Genève zoekt men naar een organisatie voor . letiging der nooden der ontwortelde vluchte- I li igen, [bannelingen, verstootenen. Men bestu- d:at ; de immigratie-mogelijkheid in landen, 1 - dij nog ruimte en werkgelegenheid voor -J-vispKKielingen hebben. Maar het gaat alles langzaam, en de economische depressie en binnehlandsch-politieke overwegingen ver- naliwen de kansen in menig land voor hen, die ' een plekje grond zoeken waar zij — elemen- tai™^verlangen! — ongemolesteerd kunnen ; levèn. En intusschen krampt de nood voor de ' groeiende scharen van „niemandsvolk". Vandaag lezen wij van Poolsche Joden in Duiuchland (en anderen, die nu in officieele Póóisöhe oogen „undesirables" zijn geworden) dife door Warschau eenvoudig met één penne- stre'efc ontnationaliseerd worden, en op het- zelfde oogenblik in officieele Duitsche oogen als voortaan reçhtelooze menschen in de ter- menf. vallen om van hun bed te worden ge- lichtf' en als „staatloozen" over de grens te worden geschoven. De grens dan van het land, dat deze menschen letterlijk en figuurlijk heeft ,, af geschreven", en hen niet meer wenscht. Dat de laatste berichten wijzen op de moge- lijkheid van eenig uitstel, verandert niets aan het wezen dezer zaak. Het is in onze twintig ste eeuw en in het beschaafde Europa moge lijk lom menschengroepen tot rechteloos vee te decreteeren (maar voor dierlijk vee zorgen ; tenminste nog de dierenbeschermers) en als vee Tin of uit het veld te jagen. Aan de j Duitsch-Tsjechische grens is iets dergelijks gebeurd. Duitschland jaagt ongewenschte ;• Joden over de nieuwe, nog niet eens afgeba- ! kende- grens. De Tsjechen weigeren om deze verdrevenen toe te laten. Nu kampeeren zij in den barren winter in een nog niet afge- ' paalde grensstrook. Niemandsvolk in nie- (mandsland! Het menscheljjke leed druipt over deze ■ zwarte bladzijde in de annalen van Europa. » Het: probleem van „niemandsvolk" verbreidt

zich over steeds meer landen. Maar de ge meenschappelijke wil om met vereende krachten een oplossing te scheppen voor iets, dat opgelost moet worden, blijft uit. En toch is zulk een oplossing een eisch van elementaire menschelijkheid. Wie uit zijn nationale midden menschengroepen verstoot (over het recht om dit te doen, spre ken wij hier nu maar met), kan met het ver stoeten niet volstaan, maar moet andere we gen wijzen. Want in onze wereld wordt met eens genoeg „niemandsland" beschikbaar ge steld om het „niemandsvolk" te herbergen, ten zij het met gezamenlijken wil en krachten wordt ge maakt en beschermd, wàår men daarvoor dan ook de mogelijkheid zal willen zoçken en vinden. Men spreekt er van, dat na München met vernieuwde kracht en vemieuwde verwach tingen moet worden gewerkt aan de regeling van een nieuwen en beteren Europeeschen vrede. Het „niemandsvolk" dat nu door ver- SchiUende landen wordt geschapen, heeft recht op een menschelijk aandeel daarin. Wat nu gebeurt, onteert Europa.

Comité voor Bijzondere Joodsche Belangen Bureaux Vluchtelingen-Comité: 's-Gravenhekje 7, Amsterdam C. Bureaux Fin. Commissie van het „Nood fonds 1933", Weesperplein 4, kamer 85J Amsterdam C, telefoon 53908 Postgiro: Nr. 214500 t.n.v. Prof. Dr. D. Cohen, Noodfonds 1933, Amsterdam. Bankier: Incasso-Bank N.V., Stadhou derskade 123, Amsterdam Z. Accountantscontrole : E. van Dien.

' /Oÿÿkf mét acßdL / UsoUajl Acoistsi , fijd tOtß. J
Z/^SOO I Å. ISoXju^, „/ 3^ '

Figure 5 – Example 2 of ALTO-XML in use: Archive from NIOD Institute https://www.archieven.nl/nl/search-modonly?mivast=298&mizig=210&miadt=298&micode=181k&milang=nl&mizk_alle=Duitsche%20vrouw&miview=inv2

3.2.3.3 Resources

- ALTO-XML is maintained by the Library of Congress. Its use is becoming more widespread with digitising companies and research projects: http://www.loc.gov/standards/alto/
- Github repository of the standard: https://github.com/altoxml

3.2.4 Music Encoding Initiative[52]

3.2.4.1 Scope

The Music Encoding Initiative (MEI) is "an open-source effort to define a system for encoding musical documents in a machine-readable structure"[53]. This encoding format, commonly (and somewhat confusingly) also referred to as MEI, is one of the many music encoding standards existing today. The format, which dates back to 1999, is based on existing encoding standards - in particular, it is modeled on the Text Encoding Initiative (TEI) specification principles, meaning that it is XML-based. It brings together researchers from various communities, including computer scientists, information scientists, musicologists, music theorists, librarians, and historians, and aims to define best practices for the representation of a broad range of musical documents and structures - thus facilitating the exchange, examination, validation and comparison of such documents. MEI is primarily catered towards an academic audience; as such, it distinguishes itself from the other XML-based music encoding format currently at the forefront, MusicXML, which has a strong commercial interest[54].

The primary reference point for researchers or others interested wanting to engage with MEI is the official website, http://www.music-encoding.org. Here one can find, among many other things, a "Gentle introduction to MEI", various more in-depth tutorials, an extensive bibliography covering the history of the project from its conception to the latest developments, the proceedings of the annual conference (see below), guidelines providing extensive documentation of the different components of the MEI model as well as best practice suggestions, and an overview of tools and projects that utilise MEI (more on tools and projects below).

The MEI community maintains an official mailing list, MEI-L, which is used as its general communication channel. Through this list, community members are informed about relevant events; moreover, it functions as a discussion platform. One such event is the annual Music Encoding Conference (MEC), which since 2013 has taken place alternately in Europe and in North America.

[52] Adeline Joffres (CNRS/Huma-Num), Reinier de Valk (DANS), Vincent Besson (CESR, MUSICA)

[53] See http://music-encoding.org.

[54] See http://www.musicxml.com.

3.2.4.2 Technical overview

The MEI format is formalised in the MEI schema[55], an XML schema that provides a set of rules for recording characteristics, both content-related and physical, of documents containing music notation. More specifically, MEI is designed to be able to distinguish between four domains that separate the functions of the symbols within a music representation system. The *logical* domain contains the musical content as provided by the composer (pitches, durations, dynamics); the *gestural* domain contains information that may be added by a performer when interpreting the logical domain (timing, phrasing); the *visual* domain contains information on the visual appearance of the score (page layout, musical font); and the *analytical* domain contains analyses of the content in any of the other three domains.

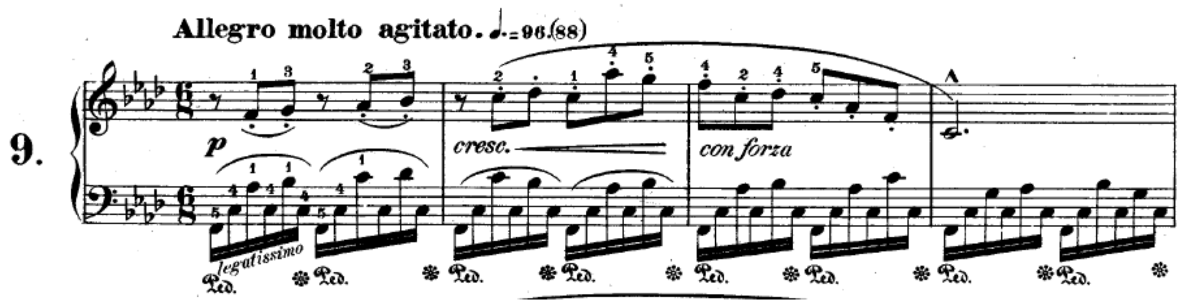


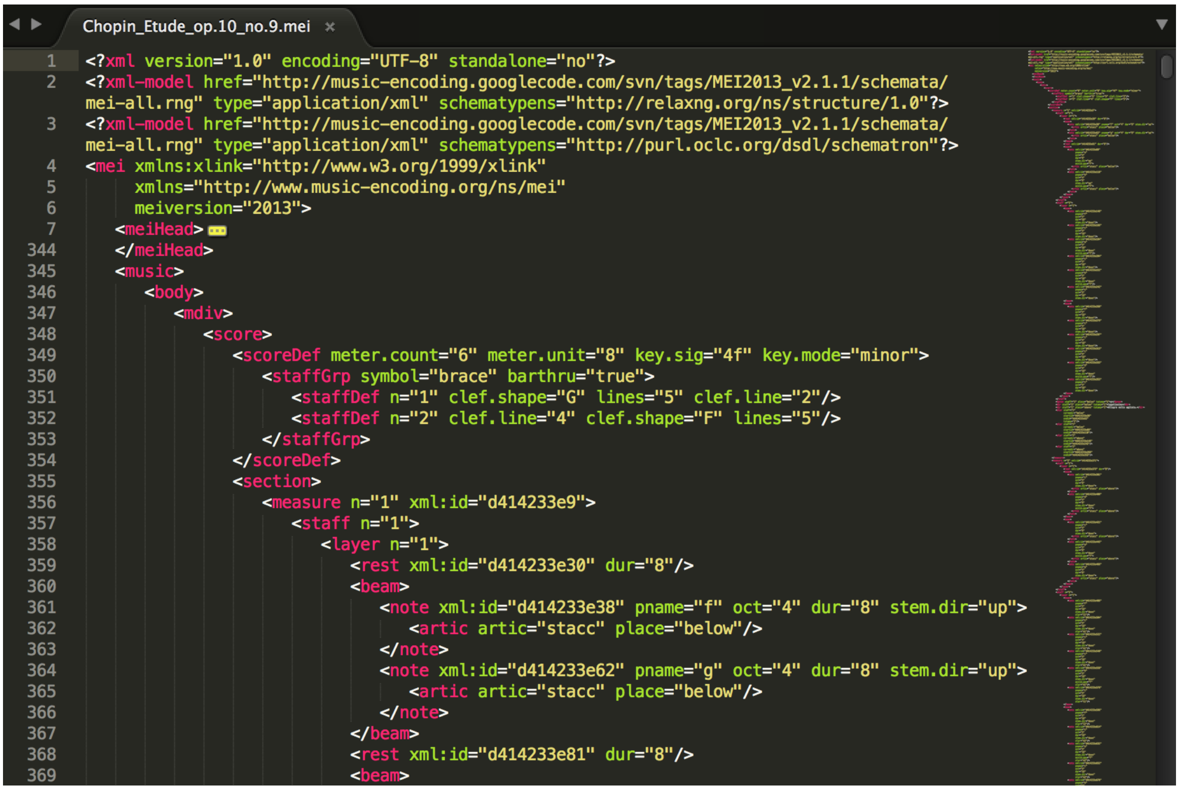
Figure 6 - Frédéric Chopin, Étude Op. 10, No. 9, opening bars

As is common with XML-based formats, each MEI file can be validated against the MEI schema to ensure that the rules set out in the schema are followed in the file[56]. A valid MEI file contains at least two sub-elements within the parent <mei> element: <meiHead>, containing metadata (composer, title, provenance, etc. - metadata can be described in great detail), and <music>, containing the information belonging to the four domains as described above. This basic structure is clearly visible in Figure 7, which shows the first lines of an MEI encoding of the musical fragment shown in Figure 6 (the <meiHead> element has been collapsed to make the figure fit the page).[57] Note that the encoding contains element attributes that belong to the logical domain (pname, oct, dur), the gestural domain (artic), and the visual domain (place, stem.dir).

[55] See http://music-encoding.org/downloads/latest-release.

[56] It is possible to alter the MEI schema to enable the encoding of "uncommon practices" (e.g., avant-garde, non-Western, or ancient forms of notation). This process is called *schema customization*; see http://music-encoding.org/support/a-gentle-introduction-to-mei (introduction) and http://music-encoding.org/tools/customization-service (customisation tool).

[57] The file was downloaded from https://github.com/music-encoding/sample-encodings/releases/tag/v3.0.0.



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<?xml-model href="http://music-encoding.googlecode.com/svn/tags/MEI2013_v2.1.1/schemata/
mei-all.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="http://music-encoding.googlecode.com/svn/tags/MEI2013_v2.1.1/schemata/
mei-all.rng" type="application/xml" schematypens="http://purl.oclc.org/dsdl/schematron"?>
<mei xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns="http://www.music-encoding.org/ns/mei"
     meiversion="2013">
   <meiHead>…
   </meiHead>
   <music>
      <body>
         <mdiv>
            <score>
               <scoreDef meter.count="6" meter.unit="8" key.sig="4f" key.mode="minor">
                  <staffGrp symbol="brace" barthru="true">
                     <staffDef n="1" clef.shape="G" lines="5" clef.line="2"/>
                     <staffDef n="2" clef.line="4" clef.shape="F" lines="5"/>
                  </staffGrp>
               </scoreDef>
               <section>
                  <measure n="1" xml:id="d414233e9">
                     <staff n="1">
                        <layer n="1">
                           <rest xml:id="d414233e30" dur="8"/>
                           <beam>
                              <note xml:id="d414233e38" pname="f" oct="4" dur="8" stem.dir="up">
                                 <artic artic="stacc" place="below"/>
                              </note>
                              <note xml:id="d414233e62" pname="g" oct="4" dur="8" stem.dir="up">
                                 <artic artic="stacc" place="below"/>
                              </note>
                           </beam>
                           <rest xml:id="d414233e81" dur="8"/>
                           <beam>
```

Figure 7 - Screenshot of an MEI encoding of the fragment shown in Figure 6

3.2.4.3 Resources

- Official website: http://music-encoding.org.
- MEI GitHub repository, containing the MEI schema (as well as various customisations), the MEI guidelines, sample encodings, stylesheets, and the source code and documentation for a number of tools: https://github.com/music-encoding.
- Most of the tools have their own GitHub repository (see above).
- Bibliography: an extensive bibliography can be found at http://music-encoding.org/community/bibliography.
- Bibliography on Zotero: https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/ZVPNPUJK

Tools

The MEI community offers a number of open source tools for working with MEI data. An overview (where available, the website and location of the source code and documentation are listed as well)[58]:

[58] All links can also be found at http://music-encoding.org/tools.

- Customization Service: an online service for customising the MEI schema (see above), either limiting or extending it.
    - http://custom.simssa.ca
    - https://github.com/music-encoding/customeization
- Sibelius to MEI Plugin: a Sibelius[59] plugin that enables export as MEI.
    - https://github.com/music-encoding/sibmei/releases
- Verovio: a music notation engraving library in C++ that can be compiled and wrapped into different programming languages.
    - http://www.verovio.org/index.xhtml
    - https://github.com/rism-ch/verovio
- MEI to music21 Converter: a Python module for the music21 toolkit[60], enabling the import of MEI files.
    - http://web.mit.edu/music21/doc/moduleReference/moduleMeiBase.html
- MEItoVexFlow: a JavaScript library that converts MEI into drawing instructions for the VexFlow online music notation rendering API[61].
    - http://tei-music-sig.github.io/MEItoVexFlow
    - https://github.com/TEI-Music-SIG/MEItoVexFlow
- LibMEI: a C++ library for reading and writing MEI files.
    - https://github.com/DDMAL/libmei/
- MEI Score Editor (MEISE): an Eclipse-based[62] music notation editor for viewing and editing MEI files.
    - https://de.dariah.eu/mei-score-editor
    - https://sourceforge.net/projects/meise
- Metadata Editor and Repository for MEI Data (MerMEId): a JavaScript library for the editing, handling, and (pre-)viewing of music metadata in MEI files.
    - http://www.kb.dk/en/nb/dcm/projekter/mermeid.html

3.2.4.4 Ongoing efforts

Projects

The MEI format is used in a considerable number of projects initiated over the past few years. A selection of projects, ordered by category, is shown overleaf [63].

[59] See http://www.avid.com/sibelius.

[60] See http://web.mit.edu/music21.

[61] See http://www.vexflow.com.

[62] See https://eclipse.org.

- Digital critical editions
    - *Gesualdo Online Project* (CESR, CNRS, Université François-Rabelais Tours)
    - *Beethovens Werkstatt* (Universität Paderborn, Beethoven-Haus Bonn, Detmold Hochschule für Musik)
    - *Edirom Project* (Musikwissenschaftliches Seminar Detmold/Paderborn)
- Repertory analysis
    - *Lost Voices Project* (Haverford College, CESR)
    - *Citations: The Renaissance Imitation Mass (CRIM)* (Maryland Institute for Technology in the Humanities*, Haverford College & CESR)*
- Metadata
    - *Catalogue of Carl Nielsen's Works (CNW)* (Danish Center for Music Editing)
- At the intersection of different fields
    - *Enhancing Music Notation Addressability* (EMA) (Maryland Institute for Technology in the Humanities)
    - *Single Interface for Music Score Searching and Analysis (SIMSSA)* (McGill University)

N*ew developments and perspectives*

Two approaches can be developed and deepened within PARTHENOS WP4 on standardization around MEI :

- To compare the scope and intended community of a number of music encoding formats. First steps in this direction have been taken; results will be presented at this year's MEC in Tours (France).
- Another topic could be to follow the current work of the MEI community on the conception of a 'MEI Lite' (the final name is yet to be decided), i.e., a simplified version of the format. The foreseen questions to go ahead with that work could be: What will MEI Lite offer? A 'light' version? An educational version? Or a simplified support for archiving? These questions are still unanswered but should offer the community some prospects for the future.

The MEI consortium is also currently developing various ideas based on specific projects to enlarge the scope and functionalities of the MEI standard that PARTHENOS could follow the progress of:

[63] A complete overview, as well as links to the individual projects, can be found at http://music-encoding.org/community/projects-users.

- The possibility to put MEI within an Omeka CMS which could integrate a MEI viewer (Verovio http://www.verovio.org/index.xhtml) : development and implementation of an interoperable tool : TiKiT•MUSICA in France (Tours, MUSICA Huma-Num's consortium[64]).
- The creation of musical nano-publications from MEI files (http://mith.umd.edu/research/enhancing-music-notation-addressability);
- The possibility of musical analysis through artificial intelligence.

3.3 Language resources

3.3.1 Part of speech tagging - morphosyntactic Annotation Framework (MAF)[65]

3.3.1.1 Overview

Morphosyntactic Annotation Framework (MAF), an ISO standard 24611:2012, is intended to provide a data model for morphosyntactic annotation of textual data, i.e. grammatical classes (part of speech, e.g. noun, adjective, verb), morphological structure of words and grammatical categories (e.g. number, gender, person). Rather than proposing a single tagset or a family of tagsets the standard offers a generic way to anchor, structure and organize annotations. The standard also describes an XML serialization for morphosyntactic annotations, with equivalences to the guidelines of the TEI (Text Encoding Initiative).

Raw original document is accompanied by a set of annotations – word forms covering a set of tokens, identifying non-empty continuous parts of the document. The material corresponding to a token can be annotated in the document itself (inline annotation) or identified by a pair of document positions (e.g. character offsets, time durations for speech, frames for video etc.) as standoff annotations.

Word forms may be associated to tokens (in a many-to-many model), may embed word form subcomponents to represent compound terms and link output of tokenization to some lexicon. Word forms provide morphosyntactic information about a word (POS, lemma, morphology etc.) by means of specifying feature structures conformant to a tagset.

[64] See: http://musica.hypotheses.org/.

[65] Maciej Ogrodniczuk (CLARIN), with modifications by Piotr Banski (CLARIN) and Laurent Romary (INRIA)

Tagset data (types, features, feature values) may be mapped to data categories from ISOCat (or equivalent) data category registry, and feature structure declarations may be used to identify valid morphosyntactic content. Similarly, feature structure libraries may be used to name the most common morphosyntactic contents.

Structural ambiguities are represented by lattices – direct acyclic graphs with single initial and terminal nodes. Lexical ambiguities can be handled by using alternations on word forms while morphological ambiguities by alternations inside feature structures.

3.3.1.2 Usage examples

Word form corresponding to agglutinated morpheme:

```
<token id="t0">aujourd</token>
<token id="t1" glue="""">hui</token>
<wordForm entry="aujourd'hui" tokens="t0 t1">
```

Contracted word forms:

```
<token id="t0">isn't</token>
<wordForm entry="is" tokens="t0">
<wordForm entry="not" tokens="t0">
```

Morphological ambiguities:

```
<wordForm entry="eat">
  <token>eat</token>
  <fs>
    <f name="pos" fVal="v"/>
    <f name="pers">
      <vAlt>
        <sym value="1"/>
        <sym value="2">
      </vAlt>
    </f>
    <f name="tense" fVal="pres"/>
    <f name="mode" fVal="ind"/>
  </fs>
</wordForm>
```

Structural variants:

```
<fsm>
  <state id="s1" type="init"/>
  <state id="s2"/>
  <state id="s3"/>
  <state id="s4" type="fina1"/>
  <transition source="s1" target="s4">
    <wordForm tokens="3 4 5" entry="potato" .../>
  </transition>
  <transition source="s1" target="s2">
```

```
    <wordForm tokens="3" entry="apple" .../>
  </transition>
  <transition source="s2" target="s3">
    <wordForm tokens="4" entry="from" .../>
  </transition>
  <transition source="s3" target="s4">
    <wordForm tokens="5" entry="earth" .../>
  </transition>
</fsm>
```

Figure 8 – Examples of MAF in use

3.3.1.3 References

- ISO 24611:2012. *Language resource management – Morpho-syntactic annotation framework* (MAF).
- Clément L., de la Clergerie É. (2005). *MAF: a morphosyntactic annotation framework*. In the Proceedings of the Second Language and Technology Conference, Poznań, Poland.
- Monachini, M., Calzolari N. (1994). *Synopsis and Comparison of Morpho-syntactic Phenomena Encoded in Lexicon and Corpora. A Common Proposal and Applications to European Languages*. Internal Document, EAGLES Lexicon Group, ILC, Università Pisa, October 1994.
- Przepiórkowski A., Bański P. (2011). *Which XML standards for multilevel corpus annotation?*
- In Z. Vetulani (ed.) Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference (LTC 2009), Poznań, Poland, November 6–8, 2009. Revised Selected Papers, vol. 6562 of Lecture Notes in Artificial Intelligence, pp. 400–411, Berlin, 2011. Springer Verlag.

3.3.1.4 Ongoing efforts

MAF has become a stable background document for anyone designing annotation schemes and tagsets in the domain of morpho-syntactic annotation. The priority should be set on defining better documentation material for the standards as well as a systematic alignment with the TEI guidelines.

3.3.2 Syntax Annotation Framework (SynAF)[66]

3.3.2.1 Overview

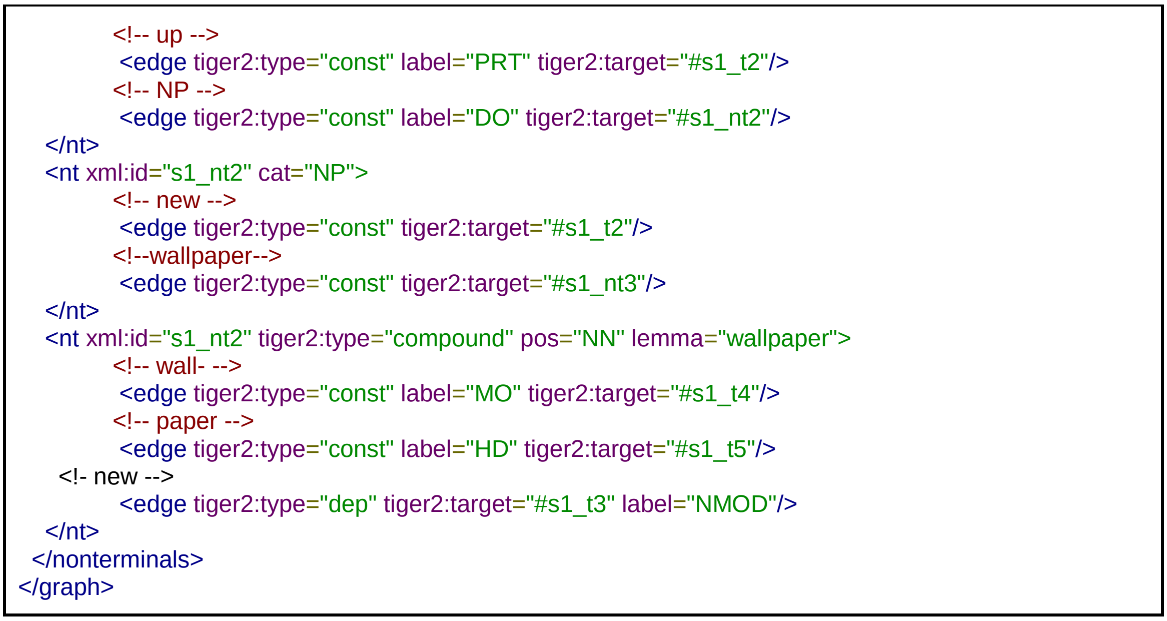
Syntactic Annotation Framework (SynAF), a multi-part ISO standard 24615:2010, is intended to represent the syntactic annotation of textual data such as grammatical features, phrase structures and dependency structures. SynAF defines both a meta-model for syntactic annotation (graphs made of nodes and edges) and a set of data categories. Syntactic nodes are either terminal nodes equivalent to MAF word forms, annotated with syntactic data categories according to the word level, or non-terminal nodes annotated with syntactic categories from the phrasal, clausal and sentential level. Relations between syntactic nodes, such as dependency or constituency relations, are represented with syntactic edges. Annotations can be applied to nodes and edges. The standard does not propose a specific tagset but only generic classes and specific data categories. Annotation vocabulary should be defined by means of a data category registry, e.g. ISOCat or equivalent. Several possible serialization formats may be used such as TIGER-XML or Graph Annotation Format defined in LAF. The example below shows a graphical representation of a multi-layer syntactic annotation in an early serialization format, Tiger2, currently part of the standardization process of ISO Tiger (the 2nd part of the SynAF specification).

1.3.2.2 Serialization example

The attributes in the "tiger2" namespace represent an overlay upon the Tiger XML format and contain references to a separate tokens.xml document in ISO MAF.

```
<graph root="s1_ROOT" discontinuous="true">
  <terminals>
    <t xml:id="s1_t1" pos="VB" lemma="put"
          tiger2:corresp="tokens.xml#wordForm1">
           <edge tiger2:type="dep" tiger2:target="#s1_nt2" label="OBJ"/>
           <edge tiger2:type="dep" tiger2:target="#s1_t2" label="PRT"/>
    </t>
    <t xml:id="s1_t2" pos="RP" lemma="up"
          tiger2:corresp="tokens.xml#wordForm2"/>
    <t xml:id="s1_t3" pos="JJ" lemma="new"
          tiger2:corresp="tokens.xml#wordForm3"/>
    <t xml:id="s1_t4" tiger2:type="stem"
          tiger2:corresp="tokens.xml#wordForm4"/>
    <t xml:id="s1_t5" tiger2:type="stem"
          tiger2:corresp="tokens.xml#wordForm5"/>
  </terminals>
  <nonterminals>
          <nt xml:id="s1_nt1" cat="VP">
          <!-- put -->
           <edge tiger2:type="const" label="HD" tiger2:target="#s1_t1"/>
```

[66] Maciej Ogrodniczuk (CLARIN), with modifications by Piotr Banski (CLARIN) and Laurent Romary (INRIA)



```
        <!-- up -->
         <edge tiger2:type="const" label="PRT" tiger2:target="#s1_t2"/>
        <!-- NP -->
         <edge tiger2:type="const" label="DO" tiger2:target="#s1_nt2"/>
  </nt>
  <nt xml:id="s1_nt2" cat="NP">
        <!-- new -->
         <edge tiger2:type="const" tiger2:target="#s1_t2"/>
        <!--wallpaper-->
         <edge tiger2:type="const" tiger2:target="#s1_nt3"/>
  </nt>
  <nt xml:id="s1_nt2" tiger2:type="compound" pos="NN" lemma="wallpaper">
        <!-- wall- -->
         <edge tiger2:type="const" label="MO" tiger2:target="#s1_t4"/>
        <!-- paper -->
         <edge tiger2:type="const" label="HD" tiger2:target="#s1_t5"/>
   <!- new -->
         <edge tiger2:type="dep" tiger2:target="#s1_t3" label="NMOD"/>
  </nt>
 </nonterminals>
</graph>
```

Figure 9 – XML representation of the Verb Phrase "put up new wallpaper" in Tiger2.

3.3.2.3 References

- ISO 24615. *Language resource management—Syntactic annotation framework (SynAF)*.
- Bunt H., Alexandersson J., Choe J.-W., Fang A. C., Hasida K, Petukhova V., Popescu-Belis A., Traum D. (2012). *ISO 24617-2: A semantically-based standard for dialogue annotation.* In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 430–437. European Language Resources Association (ELRA).
- Bunt H., Prasad R., Joshi A. (2012) *First Steps Towards an ISO Standard for Annotating Discourse Relations*. In Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools, Istanbul, Turkey, pp. 60–69. European Language Resources Association (ELRA).
- Declerck, T. (2006). *SynAF: Towards a Standard for Syntactic Annotation*. In Proceedings of LREC 2006, pp. 229–232. European Language Resources Association (ELRA).
- Pustejovsky J., Lee K., Bunt H., Romary L. (2010). *ISO-TimeML: An International Standard for Semantic annotation*. In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 394–397. European Language Resources Association (ELRA).

- Romary, L., Zeldes A., Zipser F. (2015). *<tiger2/> – Serialising the ISO SynAF Syntactic Object Model.* Lang Resources & Evaluation 49: 1. doi:10.1007/s10579-014-9288-x.
- Stührenberg M. (2012). *The TEI and Current Standards for Structuring Linguistic Data: An Overview*. Journal of the Text Encoding Initiative (3), pp. 1–14. http://jtei.revues.org/523.

3.3.2.4 Ongoing efforts

Part two of the standard, dedicated to the XML serialisation of the SynAF model, is about to be published under the name of ISO Tiger in the summer 2017 and we should take this opportunity to widely inform the community about this, in particular through the CLARIN network. The coming period should also be dedicated to the design of part 3, which should offer a TEI based serialisation for the SynAF meta-model, that should be at least as expressive as ISO Tiger.

3.3.3 Stand-off annotation in TEI[67]

3.3.3.1 Overview

Stand-off annotation assumes that the source text of the corpus, ideally kept in an unannotated form and in read-only files, is the root of independent possibly multi-file system of data descriptions (each description focusing on a distinct aspect of the source data). The source text is typically accompanied by a level of primary segmentation, which may be the lowest-level XML layer of annotation. The other files form a possibly multi-leaved and multi-leveled hierarchy referencing either the level of primary segmentation, or higher order levels of description.

For constructing a simple working stand-off-annotated corpus prototype, portions of TEI Guidelines (chapter 15 on language corpora, chapter 16 on stand-off linking and chapter 17 on analytical mechanisms) should be consulted.

3.3.3.2 Use case: Stand-off annotation in the National Corpus of Polish

The annotation architecture of the one-billion-word National Corpus of Polish (http://nkjp.pl/) follows the guidelines of the stand-off annotation to the extent allowed by the TEI schema. Each corpus text (text.xml) is kept in a separate directory together with

[67] Maciej Ogrodniczuk (CLARIN), with modifications by Piotr Banski (CLARIN) and Laurent Romary (INRIA)

the annotation files that reference it directly or indirectly (ann_structure.xml, ann_segmentation.xml, ann_morphosyntax.xml etc.), and with the header that is included by all these files (header.xml). All of these files contain TEI documents forming a hierarchy of annotation levels, as presented below:

The text.xml file is the root, referenced by the layer of text structure (providing markup from the paragraph level upwards) and the layer of segmentation. The segmentation layer is further referenced by the layer of morphosyntactic information and word-sense annotation. The morphosyntactic level, in turn, is the basis for the level identifying syntactic words, which constitutes the foundation upon which the levels identifying syntactic chunks and named entities are built.

In text.xml, the normalized source text is divided in paragraph-sized chunks (enclosed in anonymous blocks, <ab>, to be further refined in the text-structure level of annotation). It also includes two headers: the main corpus header, which encodes information relevant to all parts of the corpus, and the local header, which records the information on the particular text and its annotations.

The segmentation file provides the base segmentation level that is further used as the basis for other kinds of annotation. It is implemented as a TEI document with <seg> elements that contain references to string ranges from text source file.

The morphosyntactic layer of annotation consists of a series of elements that contain TEI feature structures (i) providing basic information on the segment, (ii) specifying the possible interpretations as identified by the morphological analyser, and (iii) pointing at the morphosyntactic description selected by the disambiguating agent.

The higher-order annotation layers also contain feature structures, which usually point at the selected segments of annotation layers that are one level lower, and identify their function within the given data structure.

Normalized source (text.xml):

```
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude"
           xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="corpus_header.xml"/>
```

```
<TEI>
 <xi:include href="header.xml"/>
 <text xml:id="txt_text" xml:lang="pl">
  <body xml:id="txt_body">
   <div xml:id="txt_1-div" decls="#h_1-bibl">
    <ab n="p1in1of:DL_1056" xml:id="txt_1.1-ab">Fakt, że...</ab>
   </div>

   ...
```

Segmentation file (ann_segmentation.xml):

```
<teiCorpus>
<TEI>
 <text xml:id="segm_text" xml:lang="pl">
  <body xml:id="segm_body">
   <p corresp="text.xml#txt_1-div" xml:id="segm_1-p">
    <s xml:id="segm_1.20-s">
     <!-- Fakt -->
     <seg corresp="text.xml#string-range(txt_1.1-ab,0,4)"
          xml:id="segm_1.1-seg"/>
                  ...
```

Morphosyntactic description (ann_morphosyntax.xml):

```
<teiCorpus>
<TEI>
 <text>
  <body>
   <p corresp="ann_segmentation.xml#segm_1-p" xml:id="morph_1-p">
    <s corresp="ann_segmentation.xml#segm_1.20-s" xml:id="morph_1.20-s">
     <seg corresp="ann_segmentation.xml#segm_1.1-seg"
             xml:id="morph_1.1-seg">
      <fs type="morph">
       <f name="orth">
        <string>Fakt</string>
       </f>
       <!-- Fakt [0,4] -->
       <f name="interps">
        <fs type="lex" xml:id="morph_1.1.1-lex">
         <f name="base">
          <string>fakt</string>
         </f>
         <f name="ctag">
          <symbol value="subst"/>
         </f>
         <f name="msd">
          <vAlt>
           <symbol value="sg:nom:m3" xml:id="morph_1.1.1.1-msd"/>
           <symbol value="sg:acc:m3" xml:id="morph_1.1.1.2-msd"/>
          </vAlt>
         </f>
        </fs>
       </f>
       <f name="disamb">
        <fs feats="#pantera" type="tool_report">
         <f fVal="#morph_1.1.1.1-msd" name="choice"/>
         <f name="interpretation">
          <string>fakt:subst:sg:nom:m3</string>
         </f>
   </fs>
  </f>
 </fs>
</seg>
        …
```

Figure 10 – TEI example of annotation of the National Corpus of Polish

3.3.3.3 References

- Bański P., Przepiórkowski A. (2009). *Stand-off TEI annotation: the case of the National Corpus of Polish*. In Proceedings of the 3rd Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009, pp. 64–67, Singapore, 2009.
- Bański P. (2010). *Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless*. In Proceedings of Balisage: The Markup Conference, 2010. 10.4242/BalisageVol5.Banski01.
- Bański, P., Wójtowicz, B. (2010). *The Open-Content Text Corpus project*. In V. Arranz., L. van Eerten (eds.) Proceedings of the LREC workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LRSLM2010), 23 May 2010, Valletta, Malta, pp. 19–25. Available from http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf.
- Banski P., Gaiffe B., Lopez P., Meoni S., Romary L., et al.. Wake up, standOff!. TEI Conference 2016, Sep 2016, Vienna, Austria. <http://tei2016.acdh.oeaw.ac.at>. <hal-01374102>
- Javier P., Lopez P. and Romary L. A Generic Formalism for Encoding Stand-off annotations in TEI. 2014. <hal-01061548>
- TEI Consortium (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.1.0. Last updated on 15th December 2016. http://www.tei-c.org/Guidelines/P5/.

3.3.3.4 Ongoing efforts

Since two years ago, various use cases and samples have been gathered on GitHub to pave the way to a first proposal introducing a new standOff element in the TEI guidelines. The ongoing period will be dedicated to the finalisation of the specification before it is officially transmitted to the TEI technical council for discussion and, hopefully, approval.

The objective is to have the proposal approved for the TEI conference that will take place in the Autumn of 2018.

3.4 Models and formats for lexical information

3.4.1 LMF diachrony[68]

3.4.1.1 Scope

The scope of this standard will cover the encoding of all lexical, conceptual and metadata relevant to born digital and retro-digitized etymological datasets. They are as follows:

- Etymological processes;
- Dating and sequence;
- Language information;
- Lexical forms; orthographic and phonetic
- Related forms: etymons, roots, cognates
- Grammatical information
- Semantic information
- Bibliographic information
- Notes: editors notes and other common miscellaneous content
- Level of confidence
- External references to ontological or other knowledge sources

3.4.1.2 Technical overview

The working contents of the LMF diachrony section are shown in the UML diagram below in conjunction with the core model. In addition to containing any content that may be relevant to or included in a typical synchronic dictionary entry (e.g. forms, sense, etc.), the 'Etymology' section can occur 0..n times and may occur recursively to express the etymological pathway of an etymon.

The section 'Chain' can be used to express a diachronic sequence of form changes, in particular with phonetic forms. Future additions will include a means to explicitly express a

[68] Jack Bowers (OEAW), Fahad Khan (CNR-ILC), Mohamed Khemakhem (INRIA), Monica Monachini (CNR-ILC), Laurent Romary (INRIA)

set of cognates in related languages (‘CognateSet’). ‘CognateSet’ can contain 1..n cognate forms.

Note: The dotted arrow from ‘Etymon’ to ‘LexicalEntry’ is formatted in that way due to the undecided nature of this particular question. Specifically, it has not yet been determined whether an Etymon should, by default, contain all the possible and required components of full lexical entry (which would necessarily be required to contain a lemma) or whether it should have its own substructure which would allow lemma to be optional.

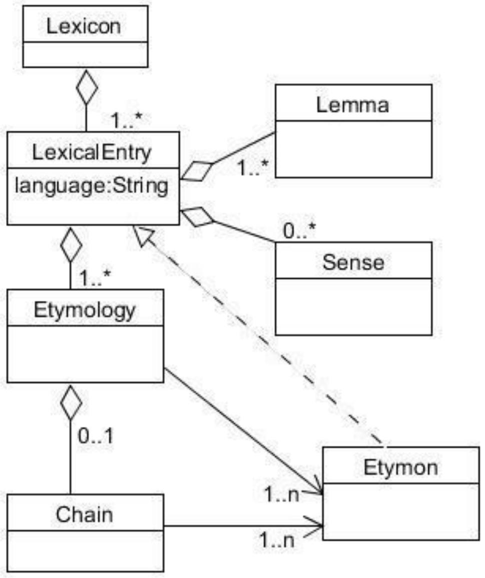


Figure 11 - Working LMF Diachrony with Core Model

3.4.1.3 Resources

Git Repository

https://github.com/anasfkhan81/LMFEty

Bibliography

Bowers, J., & Romary, L. (2016). Deep encoding of etymological information in TEI. Retrieved from https://hal.inria.fr/hal-01296498/

Salmon-Alt S., L. Romary, E. Buchi (2005). “Modeling Diachrony in Dictionaries”. ACH-ALLC 2005, Vancouver, Canada.

Salmon-Alt Susanne (2006) “Data structures for etymology: towards an etymological lexical network”, BULAG 31 1-12 — http://hal.archives-ouvertes.fr/hal-00110971

3.4.1.4 Ongoing efforts

We have identified several issues in which the previous LMF serialization lacked a means of expressing key concepts relevant to both synchronic and diachronic data, they are listed as follows:

- allow @id on any element;
- add attribute encoding dates;
- pointer attributes expressing sequence (e.g., @prev @next);
- Add <Bibliography> element
- Expand the usage of @script (equivalent to TEI @notation) to all necessary elements
- Introduce <Note> element

3.4.2 TEI Lex0 (ENeL)[69]

3.4.2.1 Scope

TEI has provided the lexicographic community with diverse alternatives for encoding different kinds of lexical resources. The flexibility that this de-facto standard ensures has engendered an explosion of the TEI schemes and consequently limited exchange and exploitation possibilities by the means of common Natural Language Processing systems.

We do not aim here to specify a mandatory format for the variety of dictionary content that we deal with, but to define a baseline encoding (TEI-Lex-0) against which existing dictionaries can be compared, and which could serve as a target transformation format for generic querying or visualization tools. Aggregating such a baseline relies on the restriction of the use of TEI elements the refinement of their definitions, and if necessary, removal of any persistent ambiguity. The outcome of such a customization would be best practice guidelines accompanied by illustrative dictionary samples.

[69] Jack Bowers (OEAW), Mohamed Khemakhem (INRIA)

3.4.2.2 Technical overview

Our starting point is the TEI P5 guidelines which is dedicated to encoding the lexical information in born digital or retro-digitized dictionary sources. To derive a restricted customization of the broader guidelines, the efforts are articulated around four major axes:

- Unified representation of lexical entry's macro structure:

A set of issues regarding the restriction of entry-like elements for a generic representation are identified which advise that elements <entryfree>, <superEntry> and <re> be replaced by a simple <entry>. The definition and organization of <entry> is reviewed to be more general and able to encode fine grained structures.

- Towards a more systematic use of sense:

Becoming mandatory, <sense> is also refined in terms of its contained elements. Whereas some elements are being recommended, such as <cit> and <usg>, others that whose function are redundant or too narrow, such as <hom> are excluded from the new scheme. A further review of the actual definition of these elements, as well as for the related elements is required.

- Recommendations for the encoding of written and spoken forms

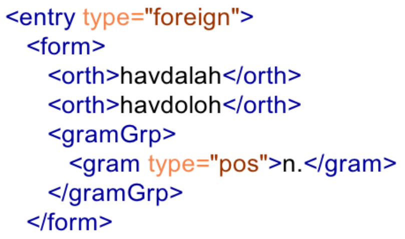
The current TEI formulation allows an extremely wide range of encoding possibilities for such information. Constraining these alternatives is enabled through the focus on revising the

  - *Grammatical properties of lexical entries:*
  - *Representation of the lemma:*
  - *Representation of inflected forms:*
  - *Paradigms:*
  - *Representation of variants:*

- Referring mechanisms between lexical entries

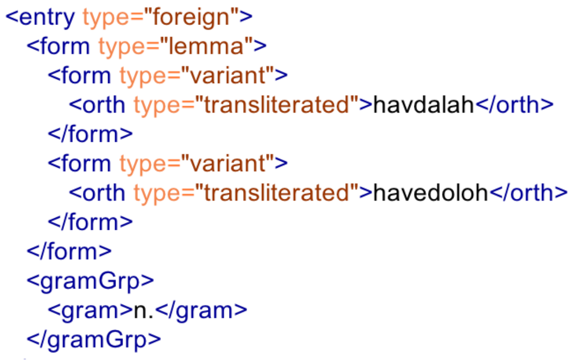
Below is an example of a the <form> portion of an entry from the TEI guidelines' Dictionary section and that same example encoded in according to TEI-Lex0. This example shows an entry of a loanword in English from Hebrew. Due to the non-standardized transliterations conventions between the two orthographies, there are two variant spellings of this word in English.

(a) Original encoding from Guidelines



```
<entry type="foreign">
  <form>
    <orth>havdalah</orth>
    <orth>havdoloh</orth>
    <gramGrp>
      <gram type="pos">n.</gram>
    </gramGrp>
  </form>
```

(b) Encoding from TEI-Lex0 section on <form>



```
<entry type="foreign">
  <form type="lemma">
    <form type="variant">
      <orth type="transliterated">havdalah</orth>
    </form>
    <form type="variant">
      <orth type="transliterated">havedoloh</orth>
    </form>
  </form>
  <gramGrp>
    <gram>n.</gram>
  </gramGrp>
```

3.4.2.3 Resources and links

Due to copyright issues, only few number of dictionary samples are made public under: https://github.com/PARTHENOSWP4/standardsLibrary/tree/master/Lexicography/ENeL-WG2

Excerpts from these dictionaries are used as a basis for our discussions and some of them will be featured in the coming TEI-Lex guidelines as illustrations of the targeted issues.

3.4.2.4 Ongoing efforts

Given the critical goal of the TEI-Lex0, several experts from different backgrounds, representing a very active community in the field of lexicography, are collaborating on this work. The labor is organized mainly in workshops held in a coordination with ENeL and DARIAH experts. These workshops consist of a sequence of parallel sessions in small groups, followed by plenary discussions. For each topic addressed during parallel sessions, the objective is to identify elements of consensus, and identify further or deeper discussion points to be addressed in the next round.

Many decisions have been already made, after the first two workshops in Berlin and Budapest. They were translated into the aforementioned working points in deep detail and after an extensive written exchange between the involved experts. For the coming meetings in Berlin and Leiden, these points are going to be refined by decisions about the remaining issues and drafting of the first official TEI-Lex-0 guidelines.

4. Standardization of reference resources

4.1 Authority lists and prosopography

4.1.1 Prosopography[70]

4.1.1.1 Scope

Prosopography[71] is the investigation of the common background characteristics of a group of actors in history, making a collective study of their lives. Prosopography is mostly used by historians to address two main research questions:

- roots of political action: e.g. the interests beneath the rhetoric of politics or the social and economic affiliations of political groupings;
- social structure and social mobility: e.g. the role in society, the degree of social mobility and the correlation of intellectual or religious movements with other factors.

Among the typical products of researchers working on prosopography there are various kinds of *repertoires*, hand lists and other reference tools, such as:

- lists of names, holders of certain offices or titles or educational qualifications;
- family genealogies;
- full biographical dictionaries, which are usually built up in part from the first two categories and in part from an infinitely wider range of resources.

With the adoption of digitisation in the humanities, traditional (printed) reference tools have been digitized, and new ones have been produced *ex-novo*: at first on CD-ROMS and DVDs and - eventually - published online. A wide range of disciplines in the Humanities and Social Sciences are represented in PARTHENOS: along with authority lists of persons and places names, a wider set of thesauri, produced in different research areas - will be available in the project content cloud. For this reason a specific VRE named RubRIcA (see infra for a detailed description) - is under development, to address all the integration needs of a complex digital research infrastructure. RubRIcA is developed in collaboration with WP2 (requirements), WP5 (modeling and mapping) and WP6 (Implementation) and will be

[70] Emilane degl'Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

[71] Stone, Lawrence, Prosopography, 1971, *Daedalus* 100:46–79.

supporting a specific use case based on integration and standardization of reference resources about prosopography.

In the supported workflow the researcher has to establish a universe to be studied, and answer a set of uniform questions (e.g. birth, death, family, social origins, economic position, place of residence, education, amount and source of personal wealth, occupation, religion, experience of office and so on). The various types of information gathered about individuals in this universe should then be compared, combined, and examined for significant variables. Finally, these types of information are tested for internal correlations and for correlations with other forms of behavior or action. At the end of the process, the researcher should be able to use the information obtained to address specific research questions (for example): make sense of political action, in order to help explain ideological or cultural change, to identify social reality and to describe and analyze with precision the structure of society and its movements.

4.1.1.2 Technical overview

There are several standards to encode prosopographical information, used in different disciplinary contexts; among the most relevant:

- EAC[72]: Encoded Archival Content. An XML schema implementing ISAAR-CPF in the Archival domain
- FOAF[73]: Friend Of A Friend. A vocabulary to provide a collection of basic terms that can be used to produce machine readable webpages for people, groups, companies etc.

A number of national authorities are also available:

- PND (now GND[74]): *Personennamendatei* is an authority file of people (for each person there is a record with: name, date of birth, occupation and PND number), built between 1995 and 1998 by German National Library and used until 2012 to provide to access to literature in libraries. PND is comparable with the Library of Congress Name Authority File (LCNAF) and since April 2012 has been integrated into the or GND (Gemeinsame Normdatei)
- Rameau: Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié

[72] http://eac.staatsbibliothek-berlin.de/index.php
[73] http://www.foaf-project.org/
[74] http://www.dnb.de/EN/Standardisierung/GND/gnd.html

4.1.1.3 Resources

- http://www.sismelfirenze.it/index.php/banche-dati/bibliotheca-scriptorum
- http://www.sismelfirenze.it/index.php/banche-dati/compendium-auctorum
- https://viaf.org/
- http://www.getty.edu/research/tools/vocabularies/ulan/index.html
- http://rameau.bnf.fr/utilisation/liste.htm

4.1.2 EAC-CPF[75]

4.1.2.1 Scope

The need for a standard structure for the recording and exchange of information about the creators of archival (and - possibly - other kind of) materials has been expressed by researchers for long a time. A group of archivists has defined the model "Encoded Archival Context - Corporate Bodies, Persons, and Families" (EAC-CPF), emphasizing its important role in archival description and its relationship with the Encoded Archival Description standard.

This standard would provide a communication standard for the exchange of authority records based on International Standard for Archival Authority Records—Corporate Bodies, Persons, Families (ISAAR(CPF)) and would parallel the standard for encoding archival record finding aids that was found in Encoded Archival Description (EAD). A separate standard would pave the way to eliminating some practical problems found in the use of EAD, which had been developed as a comprehensive solution for encoding standalone finding aids which held all forms of descriptive data about archival records.

4.1.2.2 Technical overview

The schema was submitted to SAA's Council for consideration and was fully adopted by SAA in January 2011. At that time, the EAC Working Group was disbanded and the Standards Committee of SAA formed the Technical Subcommittee for EAC-CPF, responsible for the maintenance and development of the standard going forward.

To meet the need for complexity of the CPF entities (for example, one EAC entity can represent multiple identities, or a single identity can be associated with several different

[75] Emilane degl'Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

EAC entities), the structure of a standard is necessary to account for the various ways in which EAC entities can be expressed. In order to accommodate the variety of EAC entities, EAC-CPF has adopted the following definitions:

- SINGLE IDENTITY: one person (or corporate body or family) with a single identity represented in one EAC-CPF instance. This is the most common identity type.
- MULTIPLE IDENTITY-MANY IN ONE: two or more identities (including official identities) each represented by distinct descriptions within one EAC-CPF instance. Can be programmatically converted into Multiple Identity-One in Many.
- MULTIPLE IDENTITY-ONE IN MANY: two or more identities (including official identities) each represented in two or more interrelated EAC-CPF instances. Can be programmatically converted into Multiple Identity-Many in One.
- COLLABORATIVE IDENTITY: a single identity shared by two or more persons (e.g. a shared pseudonym used in creation of a collaborative work). Use Multiple Identity-One in Many.
- ALTERNATIVE SET: derived EAC-CPF instance that is based on and incorporates two or more alternative EAC-CPF instances for the same entity. To be used by a consortium or a utility providing union access to authority records maintained in two or more systems by two or more agencies. Alternative EAC-CPF instances may be in different languages or in the same language.
- EAC-CPF has been created to accommodate this variety of identities, and includes a number of ways to express complexities based on individual repository or aggregator preferences.

These options reflect a design principle that underpins the increased opportunity for repositories or aggregators to customize the standard for specific needs while at the same time ensuring future aggregation. These flexibilities also reflect an acknowledgement that some fundamental philosophical differences, with regard to the processing of information related to separate identities of the same EAC entity, exist in the international community. EAC-CPF maintains a neutral stance on those philosophical differences, and instead has accommodated the various options without precluding aggregation in the future.
For purposes of this tag library, agents refer to repositories or services creating or maintaining EAC-CPF records, while entities refer to those for which the records are about.

4.1.3.3 Resources

- Data repository
  http://eac.staatsbibliothek-berlin.de/
- Data Schema
  http://eac.staatsbibliothek-berlin.de/eac-cpf-schemas.html
- Schema Diagram: http://eac.staatsbibliothek-berlin.de/Diagram/cpf.html
- Example Person:
  http://eac.staatsbibliothek-berlin.de/fileadmin/user_upload/schema/FRAF_P_00006.xml
- Publications
  http://eac.staatsbibliothek-berlin.de/tag-library/publications.html

4.1.3 Omeka plugin: management of authority files[76]

4.1.3.1 Scope

The solution described here aims at simplifying access to, management and interoperability of prosopographical data: it consists of a file management tool using the CMS Omeka for scholarly content, digital collection and exhibits[77], and can be handled without a steep learning curve.

This system is able to ingest and produce authority files in different formats (XML, HTML, CSV, etc) supporting different standards (Dublin Core, FOAF, TEI, EAC-CPF, etc.) without requiring any special operation from the users.

The authority records are ingested in XML markup following EAC-CPF (Encoded Archival Context - Corporate bodies, Persons and Families) convention, a quite complete format that allows to structure communities descriptions, individuals or families. It follows the indications of the second edition of ISAAR (CPF), the international standard for the description of archival producers[78].

[76] Graziella Pastore (INRIA)

[77] Omeka is developed by the Center for History and New Media at George Mason University (CHNM) https://omeka.org/. The latest version, Omeka 2.4, was released on January 21, 2016.

[78] The EAC-CPF standard is maintained by the Society of American Archivists in partnership with the Berlin State Library http://eac.staatsbibliothek-berlin.de/

4.1.3.2 Technical overview

The prosopographical corpus

The corpus of authority files EAC-CPF/XML is based on officials and jurisconsults cited in “Li livres de jostice et de plet”, a legal compilation in Old French, around 1260, taken from the mostly unpublished manuscript Paris, BnF, français 2844[79]. The prosopographical corpus is available on a GitHub repository[80].

Omeka configuration

After installing the latest Omeka version, the system must be properly configured; this solution was tested on both 2.3 and 2.4 Omeka versions.

First, it is necessary to install the following Omeka extensions: *CSV Import+* (version 2.2 improved by Daniel Berthereau)[81], *Dublin Core Extended* (version 2.0.1 by Roy Rosenzweig Center for History and New Media)[82], *XML Import* (Version 2.15 by Daniel Berthereau)[83], and *ExportEacCpf* (Version 0.1 by Graziella Pastore and Luca Foppiano)[84], the extension created for this project, which will be discussed below.

It is also necessary to define a new particular type of item, with associated metadata, in order to manage EAC-CPF elements. Omeka actually provides default item types (Text, Moving Image, Oral History, Sound, Person, etc.) to describe an item and to easily add a new element type[85]. A new item type named “Person EAC-CPF” has been created to entail a choice of EAC-CPF grouping elements, as: NameEntryParallel, existDates, places, functions, biogHist, sources, relations[86]. “Person EAC-CPF” is used in combination with Dublin Core Metadata set in order to map metadata during the importation of prosopographical files, from EAC-CPF/XML authority files into an Omeka database (basically, Dublin Core Metadata maps elements of EAC-CPF <control>, and “Person EAC-CPF” maps all others elements).

[79] http://elec.enc.sorbonne.fr/josticeetplet/

[80] https://github.com/sgraziella/prosopography_LJP

[81] https://github.com/Daniel-KM/CsvImportPlus

[82] https://omeka.org/add-ons/plugins/dublin-core-extended/

[83] https://github.com/Daniel-KM/XmlImport

[84] https://github.com/sgraziella/ExportEacCpf

[85] https://omeka.org/codex/Managing_Item_Types_2.0

[86] A more detailed description is available in the Readme.md of EacCpfExport plugin https://github.com/sgraziella/ExportEacCpf

XSL transformations and Omeka importing data

EAC-CPF is a complex XML schema and it can not easily be converted to CSV format by *XML Import* and *CSV Import* plugins. Consequently, before importing XML file and applying CSV transformations, each authority file has to be first transformed by a preliminary XSLT stylesheet in order to extract data from the original file and create a flat and more simply XML schema. This preliminary transformation could be added to the XSLT folder of *XML Import* plugin and then be recalled by the user interface, or done beforehand with dedicated software.

After importing data, it is easy to notice that Omeka lends itself only partially in XML structured content management. Actually, if Omeka interface for items management permits to duplicate input fields, it does not easily permit fitting differents fields into others; as a result, it seems complicated to reproduce the structure of each <chronItem> of EAC-CPF <biogHist> for example:

```
<biogHist>
   <chronItem>
      <date/>
      <event/>
   </chronItem>
   <chronItem>
      <dateRange>
          <fromDate/>
          <toDate/>
      </dateRange>
      <event/>
   </chronItem>
</bioghHist>
```

In the absence of a simple way to manage hierarchical elements by Omeka[87], gathering all the information present in each container element, like <chronItem>, seems to be the best choice. Following this path, the XSL transformation creates repeatable elements (called <CHRONITEM>) gathering together all the information concerning a single event.

[87] The idea to simulate hierarchical elements by delimiting them with a colon ("For example, in the full Dublin Core list of terms, "created" is a refinement of "date," so an element with the name "date:created" is sufficient" https://omeka.org/codex/Creating_an_Element_Set) seems not a good solution for a complex schema like EAC-CPF/XML, at least as regards the final display of data records. We must also draw attention to others' suggestions, as https://omeka.org/codex/Creating_an_Element_Set and http://omeka.org/forums-legacy/topic/element-set-example or to a specific implementation of Simple Vocab plugin http://omeka.org/codex/Plugins/SimpleVocab_2.0.

```
<xsl:template match="eac:biogHist">
<BIOGHIST>
      <xsl:for-each select="eac:chronList/eac:chronItem">
        <CHRONITEM>
          <xsl:value-of select="eac:dateRange"/>
          <xsl:value-of select="eac:date"/>
          <xsl:text> - </xsl:text>
          <xsl:value-of select="eac:event"/>
        </CHRONITEM>
      </xsl:for-each>
 </BIOGHIST>
</xsl:template>
```

This solutions implies a loss of information. However, the loss of the structure may be circumvented by providing typographical rules for sub-elements. For instance, punctuation elements can be introduced to differentiate internal components, before applying regular expressions in the output phase.

4.1.3.3 Resources

Github

The ExportEacCpf plugin

Omeka supports multiple output formats (omeka-json, omeka-xml, dcmes-xml, atom, METS, etc.) and new plugins can add their own custom output formats. *ExportEacCpf* is a specific plugin for Omeka to export data record from Omeka database as basic EAC-CPF/XML; this new plugin is released on a GitHub repository https://github.com/sgraziella/ExportEacCpf.

Use case

https://github.com/sgraziella/prosopography_LJP

Bibliography (Zotero)

OMEKA folder:
https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/I9X3MUTP

EAC folder:
https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/Z3ABBMDH

4.1.3.4 Ongoing efforts

This solution is ready for improvements and new openings, more or less deep and elaborate, depending on the project for which it will be used. Even if it only focuses on the EAC-CPF standard, this plugin can easily be adapted to other formats, including TEI, according to the workflow described above.

4.2 Controlled vocabularies

4.2.1 Multilingual Thesaurus Building[88]

4.2.1.1 Scope

Information resources may be of very different kinds: books, chapters in books, papers in periodicals and conference volumes, newspapers, case records, data tables, graphs, images, maps, music sheets, etc. The contents may be in different languages. These resources may be available in their conventional physical document forms and/or in digital form.

Directories, indexes, lists, catalogues and such other tools are used to discover contents and retrieve information. KOTs (Knowledge Organising Tools) are useful for managing the vocabulary/terminology of these tools. The KOTs include ontologies, taxonomies, lexicons, dictionaries, schemes for subject classifications, thesauri, wordnets, semantic nets, self-organising systems, etc. These tools are useful in order to standardise and manage vocabularies in indexes.

In a multilingual indexing thesaurus both the terms and the relationships are represented in more than one language. Since the drawing up of the Guidelines for the Establishment and Development of Multilingual Thesauri in 1976, the multilingual access to information has followed two main developments: the building of nonsymmetrical thesauri and the linking of two or more thesauri and/or controlled vocabularies.

4.2.1.2 Technical overview

There are three approaches in the development of multilingual thesauri:

[88] Emilane degl'Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

1. Building a new thesaurus from the bottom up.
    a. starting with one language and adding another language or languages
    b. starting with more than one language simultaneously

2. Combining existing thesauri.
    a. merging two or more existing thesauri into one new (multilingual) thesaurus to be used in indexing and retrieval.
    b. linking existing thesauri and subject heading lists to each other; using the existing thesauri and/or subject heading lists both in indexing and retrieval

3. Translating a thesaurus into one or more other languages.

In the last case the languages involved are not treated equally. The language of the existing thesaurus becomes the dominant language.

Linking is typically used in situations where different agencies are using their own indexing vocabularies in their own languages for their own information systems. The linking makes it possible for the end-user to search in all linked indexing vocabularies using any one of the linked thesauri or subject heading lists. An example of a multilingual linking project is the MACS project.[89]

Building from the bottom up is only viable in cases where a new thesaurus or subject heading list is envisaged. The main advantage is that the languages involved can be treated equally.

In both approaches two groups of problems are encountered:

a) Equivalence problems
Semantic problems pertain to equivalence relations between preferred and non-preferred terms in thesauri or subject heading lists. Equivalence relations exist not only within each separate language involved (intra-language equivalence), but also between the languages (inter-language equivalence).

[89] https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9041

Intra-language homonymy and inter-language homonymy are also considered semantic issues. Additional problems pertaining to semantics involve the scope, form and choice of thesaurus terms.

b) Structural problems

Structural problems involve hierarchical and associative relations between the terms. An important question in this respect is whether the structure should be the same or different for each language. In most, if not all, cases of linking, the structure will most probably not be the same in all the indexing vocabularies involved. In other approaches mentioned, it is possible in principle to apply the same structure to all languages.[90]

4.2.1.3 Resources

- https://www.ifla.org/publications/ifla-professional-reports-115
- UNESCO, *Guidelines for the Establishment and Development of Multilingual Thesauri,* Paris 1976.
- https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9041
- https://www.iso.org/standard/53657.html
- IFLA Working Group on Guidelines for Multilingual Thesauri. 2009 Guidelines for multilingual thesauri. The Hague. International Federation of Library Associations and Institutions. Available at: http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf (06 April 2017).
- International Organization for Standardization 2011 ISO 25964-1:2011, information and documentation. Thesauri and interoperability with other vocabularies. Part 1: thesauri for information retrieval. Geneva. International Organization for Standardization.
- International Organization for Standardization 2013 ISO 25964-2:2013, information and documentation. Thesauri and interoperability with other vocabularies. Part 2: interoperability with other vocabularies. Geneva. International Organization for Standardization

[90] https://www.ifla.org/publications/ifla-professional-reports-115

4.2.1.4 Ongoing efforts

Building a PARTHENOS Backbone thesaurus to homogenize the different reference materials gathered by partners using the DARIAH-ERIC BackBone Thesaurus. This activity is carried on jointly with WP5 and will be producing a VRE for Thesauri Integration.

4.2.2 BackBone Thesaurus (BBT)[91]

4.2.2.1 Scope

The BackBone Thesaurus has been undertaken by the Thesaurus Maintenance WG that was established in 2014 in the framework of DARIAH EU: The Digital Research Infrastructure for the Arts and Humanities - a research infrastructure. The BackBone Thesaurus develop a model for the interoperability between existing thesauri and ontologies.

The target of the Thesaurus Maintenance WG is to establish a coherent and integral thesaurus for arts and humanities: a “metathesaurus” that aligns all the domain specific vocabularies and terminology, therefore the BackBone Thesaurus is based on some top level entity (facets) that will become its common foundation. Using narrow terms from different thesauri for the development of top-level concepts, the BackBone Thesaurus allows experts from different sub-disciplines to align their terms by themselves under these concepts, providing a comprehensive first-level integration of terminologies, fostering a shared good practice of terminology definition and enabling cross disciplinary resource discovery, and detection of common principles.

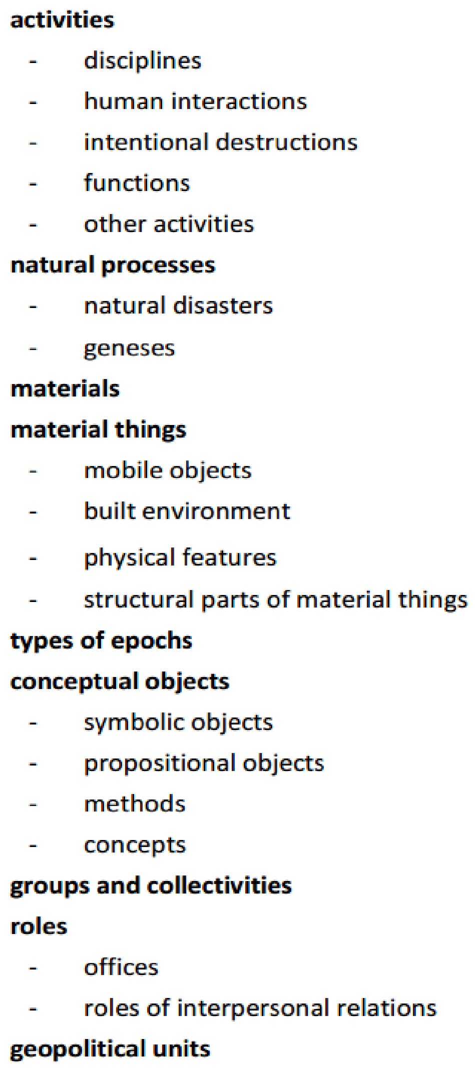
Starting from the vocabularies the Working Group had access to, an initial set of top-level concepts have been defined that constitute a first operational draft. This approach preserves backwards compatibility of new versions.

Nine facets along with their hierarchies, top terms and narrower terms’ examples have been defined thus far. Facets are the most general concepts whose properties are inherited by all possible hierarchies. The facets are further subdivided into an open number of hierarchies (expressed by the hierarchy top terms). Each Facet and hierarchy is a set or containers of terms and correspond to a top term representing the most general category all terms in this sets are narrower terms of.

[91] Emilane degl’Innocenti (CNR-OVI), Roberta Giacomi (SISMEL), Maurizio Sanesi (SISMEL)

4.2.2.2 Technical overview

The following Table 1 is a compact presentation of the terms and the top terms of the facets and the hierarchies of the backbone thesaurus (Dariah BBT). The hierarchical scheme presented has the following format: facet top terms are given in bold, hierarchy top terms are given in plain text.



activities
- disciplines
- human interactions
- intentional destructions
- functions
- other activities

natural processes
- natural disasters
- geneses

materials

material things
- mobile objects
- built environment
- physical features
- structural parts of material things

types of epochs

conceptual objects
- symbolic objects
- propositional objects
- methods
- concepts

groups and collectivities

roles
- offices
- roles of interpersonal relations

geopolitical units

4.2.2.3 Resources

- Introduction to BBT
  http://83.212.168.219/DariahCrete/en/bbt_intro_en
- BBT documentation
  http://83.212.168.219/DariahCrete/en/documents
- BBT releases
  http://83.212.168.219/DariahCrete/en/bbt_releases

4.2.3 TBX in TEI[92]

4.2.3.1 Scope

TEI offers a plethora of means for modeling lexical data. Nevertheless, those means are rooted in a semasiological approach, in which the lemma is the basis of the representation. Contrasting and complementing this view, an onomasiological approach puts the respective concepts of lexical units at its centre, i.e. all synonymous words - and in particular spanning over various languages - as associated with their concept. Such models are the basis for thesauri, synonym dictionaries, and terminological dictionaries which are commonly used in translation work, language learning, and technical writing as well as in software environments that include indexing, documentary system, or machine translation capabilities.

The present work is an adaptation of ISO standard 320042 (TBX — TermBase eXchange) and optimises the re-use of TEI constructs in combination with TBX elements. TBX is itself an application of ISO standard 16642 (TMF — Terminology Markup Framework) which provides a meta-model for the description of terminologies and other onomasiological structures. Historically, TMF has its roots in the TEI but following its fork was not able to profit from a large body of work done in the context of TEI and vice versa, the TEI lack a native model for conceptually structured lexical data. The present work is trying to bridge this gap.

4.2.3.2 Technical overview

A terminological entry is organised, following the principles of TMF, as a three level representation:

- The Terminological Entry level that represents a concept within a given subject field.
- The Language Section level that groups together all terminological descriptions for a specific language.
- The Term Section that contains all information related to a given term, comprising its graphical or phonetic representations.

The following example shows a monolingual excerpt from the TaDiRAH taxonomy in traditional TBX form. The complete model also includes its French, German, and Spanish

[92] Stefan Pernes (INRIA)

translations inside their respective <langSet> elements. In the spirit of leveraging more expressive elements from the TEI, some elements may be subject to change.

```
<termEntry xml:id="c1" xmlns="http://www.tbx.org">
  <descrip type="subordinateConceptGeneric" target="#c2">Capture</descrip>
  <descrip type="subordinateConceptGeneric" target="#c10">Creation</descrip>
  <descrip type="subordinateConceptGeneric" target="#c16">Enrichment</descrip>
  <descrip type="subordinateConceptGeneric" target="#c20">Analysis</descrip>
  <descrip type="subordinateConceptGeneric" target="#c28">Interpretation</descrip>
  <descrip type="subordinateConceptGeneric" target="#c32">Storage</descrip>
  <descrip type="subordinateConceptGeneric" target="#c37">Dissemination</descrip>
  <descrip type="subordinateConceptGeneric" target="#c44">Meta-Activities</descrip>
  <langSet xml:lang="en">
    <descrip type="definition">Research activities are usually applied to one or several research objects. An article about modeling of manuscript properties would therefore be tagged with the tags "Modeling" and "Manuscript". A plain text editor would be tagged with the tags "Writing" and "Code" and "Text".</descrip>
    <tig>
      <term>Research Activities</term>
    </tig>
  </langSet>
</termEntry>

<termEntry xml:id="c2" xmlns="http://www.tbx.org">
  <descrip type="superordinateConceptGeneric" target="#c1">Research Activities</descrip>
  <descrip type="subordinateConceptGeneric" target="#c3">Conversion</descrip>
  <descrip type="subordinateConceptGeneric" target="#c4">DataRecognition</descrip>
  <descrip type="subordinateConceptGeneric" target="#c5">Discovering</descrip>
  <descrip type="subordinateConceptGeneric" target="#c6">Gathering</descrip>
  <descrip type="subordinateConceptGeneric" target="#c7">Imaging</descrip>
  <descrip type="subordinateConceptGeneric" target="#c8">Recording</descrip>
  <descrip type="subordinateConceptGeneric" target="#c9">Transcription</descrip>
  <langSet xml:lang="en">
    <descrip type="definition">Capture generally refers to the activity of creating digital surrogates of existing cultural artefacts, or expressing existing artifacts in a digital representation (digitization). This could be a manual process (as in Transcribing) or an automated procedure (as in Imaging or DataRecognition). Such capture precedes Enrichment and Analysis, at least from a systematic point of view, if not in practice.</descrip>
    <tig>
      <term>Capture</term>
    </tig>
  </langSet>
</termEntry>
```

4.2.3.3 Resources

- Github: https://github.com/PARTHENOSWP4/standardsLibrary/tree/master/terminology
- Data repository: https://github.com/PARTHENOSWP4/standardsLibrary/tree/master/terminology/use_cases
- Bibliography (Zotero): https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/5IQ9TPWS

4.2.3.4 Ongoing efforts

1. Modeling use cases: The described approach is a testament to the fact that terminology standards are moving from an industry-driven project to a more open form, but compared to the rather clear-cut affordances of term banks, one first needs to analyse various types

of use cases specific to the TEI user community. More precise insights should be gained to identify which data categories are actually needed. For now, the use cases consist of:

- Taxonomies and ontologies specific to the humanities such as the TaDiRAH (Taxonomy of Digital Research Activities in the Humanities) and NeMO (NeDiMAH Methods Ontology)
- "Thingographies" (partly historical) such as ornithology field books, plantation account books, agricultural diaries, and punk rock fanzines
- Historical encyclopaedias such as the Brockhaus (1809), Meyer's encyclopaedic lexicon (1905), and Lemery's lexicon of chemical materials (1721)

2. Building an ODD file and adding a chapter to the TEI Guidelines: Re-introducing a native form of onomasiological data representation, but with an expanded set of elements and attributes based on the degree of expressivity in the dictionary module. Specifically, that means the TEI architecture takes priority and TEI elements will be used, where they exist. Nevertheless, it could be possible to provide a legacy/mainstream-TBX variant to such an updated <termEntry> structure.

5. Definition of standardized protocols and procedures

The resources presented in this section differ from the previous ones, as the domains covered here cannot yet rely on established standards. Therefore, this section presents roadmaps of protocols to be standardized, mostly in the domain of Cultural Heritage science.

5.1 Digital 3D objects in Arts and Humanities[93]

5.1.1 Scope

The PARTHENOS project aims to lay the foundations of a comprehensive environment revolved around the researchers' practices on and with 3D digital objects, by publishing a *White paper on "Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation"*[94]. This White paper gathers contributions from more than 25 experts of 3D imaging, modeling and processing, as well as professionals concerned with interoperability and sustainability of research data.

The topics addressed in the document are meant to help ensuring the development of standardized good practices relating to the production, the handling, the long-term conservation and the reuse of 3D objects. Therefore, even if the focus is on technical questions (formats, processing, annotation), the White Paper also points the need to clarify the legal status of 3D objects, in order to facilitate their reuse(s) in non-research contexts, in particular in Museums.

Today, the digital model has become essential for scientific documentation and analysis. However, with the rapid development and spread of 3D technology, there is an urgent need to integrate and customize the related visualization and analysis tools to support the specific needs of users within the Arts and Humanities research communities. Since the number of models produced increases exponentially, the need for efficient archival systems able to provide effective search and retrieval functionalities is also arising.

This White Paper is the result of a workshop organized by CNR (Italy), CNRS (France) and Inria (France) with support from the technical partners and on behalf of the PARTHENOS research infrastructure. It was held in Bordeaux (France), from November

[93] Adeline Joffres (CNRS/Huma-Num), Marie Puren (INRIA), Charles Riondet (INRIA)

[94] Link towards Hal

30th to December 2nd, 2016, and entitled "**Digital 3D objects in Art and Humanities: challenges of creation, interoperability and preservation**". The workshop was also supported by the work of Huma-Num's 3D-SHS consortium.

The workshop has been attended by selected PARTHENOS partners as well as some external experts, representative of both the technological and humanities domains. It aimed to enrich technical knowledge about 3D models, standards and tools in the PARTHENOS framework, addressing the common issues and epistemological questions related to the creation, use, reuse and preservation of 3D models.

More precisely, the objectives were to:

- Identify best practices and standards to ensure interoperability and sustainability;
- Expand knowledge for scholars and researchers to support 3D projects in arts, social science and humanities;
- Bridge the gap between technical people and humanities scholars (contributing to a better understanding of technologies potential and user needs);
- Share general and targeted knowledge on 3D objects issues in Art and Humanities;
- Contribute to best practices in the digitization domain for archaeologists and human sciences scholars (including 3D preservation issues: representation schemas, viewers, etc).

We selected four main topics to focus on during the workshop, corresponding to the life cycle and the various uses of 3D objects in the Humanities: (a) production and processing, (b) visualization and analysis, (c) description and preservation, and (d) bridges between Cultural Heritage and Museology. For each one of these, a number of sub-topics and issues were discussed by domain specialists in brief presentations followed by a free discussion. Those topics are the basis of the core chapters of this white paper.

In this, we intended to provide a framework for the current status of technologies, the needs and perception of DH scholars/users, and a glimpse of the near future (how can we consolidate and extend technologies by the use of standardised practices? How could we use them in an innovative manner to solve DH problems?).

The goal is to assess the needs and potentialities beyond the PARTHENOS community and to ensure that the results of the discussion will not be biased by the background of the project participants. While the reference domain is digital humanities and archaeology, we also aimed at including all related domains, such as museology, or cultural heritage at large.

We report here the results of the discussion at the workshop, further improved and extended by subsequent work done after the workshop by the participants involved. Our aim with this white paper is to briefly review the status of the technologies concerning digital 3D objects, highlighting current issues and potential for the application of those technologies in the Digital Humanities domain. Some suggestions on potential activities which could be planned and implemented in the framework of the PARTHENOS project are also presented at the end of each core section.

5.1.2 Resources

5.1.2.1 Production and processing

3D digitization for the Humanities: an overview[95]

- Bernardini, F., Rushmeier, H., 2002. The 3D Model Acquisition Pipeline. Comput. Graph. Forum 21, 149–172. doi:10.1111/1467-8659.00574
- Blais, F., 2004. Review of 20 years of range sensor development. J. Electron. Imaging 13, 231. doi:10.1117/1.1631921
- Guidi, G., Malik, U.S., 2017. Best Practices and Metrological Issues in Massive 3D Digitization of Sculptures, in: CAA 2017. Atlanta (GA), USA.
- Guidi, G., Remondino, F., 2012. 3D Modelling from Real Data, in: Modeling and Simulation in Engineering. pp. 69–102.
- Keypoints, S., Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60, 91–110. doi:10.1023/B:VISI.0000029664.99615.94
- Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., Fulk, D., 2000. The Digital Michelangelo Project: 3D Scanning of Large Statues, in: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’00. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 131–144. doi:10.1145/344779.344849

[95] References compiled by Gabriele Guidi (Politecnico di Milano, Italy)

- Ranzuglia, G., Callieri, M., Dellepiane, M., Cignoni, P., Scopigno, R., 2013. MeshLab as a complete tool for the integration of photos and color with high resolution 3D geometry data, in: CAA 2012 Conference Proceedings. Pallas Publications - Amsterdam University Press (AUP), pp. 406–416.
- Shan, J., Toth, C.K., 2008. Topographic Laser Ranging and Scanning: Principles and Processing. CRC Press.

Photogrammetric acquisition: issues and trends[96]

- Boochs, F., Bentkowska-Kafel, A., Degrigny, C., Karaszewski, M., Karmacharya, A., Kato, Z., ... & Tamas, L. (2014, November). Colour and space in cultural heritage: Key questions in 3D optical documentation of material culture for conservation, study and preservation. In Euro-Mediterranean Conference (pp. 11-24). Springer International Publishing.
- Dellepiane, M., Cavarretta, E., Cignoni, P., & Scopigno, R. (2013, June). Assisted multi-view stereo reconstruction. In 3DTV-Conference, 2013 International Conference on (pp. 318-325). IEEE.
- Pamart, A., Guillon, O., Vallet, J. M., & Luca, L. D. (2016). Toward a Multimodal Photogrammetric Acquisition and Processing Methodology for Monitoring Conservation and Restoration Studies.
- Remondino, F., Spera, M. G., Nocerino, E., Menna, F., & Nex, F. (2014). State of the art in high density image matching. The Photogrammetric Record, 29(146), 144-166.

Tools supporting annotation on 3D models[97]

- Apollonio F.I. et al, A 3D-centered Information System for the documentation of a complex restoration intervention, Submitted paper, 2017 (http://vcg.isti.cnr.it/activities/nettuno/)
- Attene M., Robbiano F., Spagnuolo M., Falcidieno B., Part-based Annotation of Virtual 3D Shapes, 2009
- De Luca, L., Busayarat, C., Stefani, C., Véron, P., & Florenzano, M. (2011). A semantic-based platform for the digital analysis of architectural heritage. Computers & Graphics, 35(2), 227-241.

[96] References compiled by Anthony Pamart (MAP Lab., CNRS, France)

[97] References compiled by Adeline Manuel (CNRS, MAP Lab France) and Roberto. Scopigno (CNR, ISTI, Italy)

- Havemann S., Settgast V., Berndt R., Eide O. and Fellner D.W. The Arrigo Showcase Reloaded – Towards a sustainable link between 3D and semantics, 2008
- Hunter J., Yu C., Assessing the Value of Semantic Annotation Services for 3D Museum Artefacts, 2011
- Manuel A., Véron P., De Luca L., 2D/3D Semantic Annotation of Spatialized Images for the Documentation and Analysis of Cultural Heritage, proceedings of EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2016).
- Ponchio F., Dellepiane M., Multiresolution and fast decompression for optimal web-based rendering , Graphical Models, Volume 88, page 1-11, 2016
- Potenziani M., Callieri M., Dellepiane M., Corsini M., Ponchio F., Scopigno R., 3DHOP: 3D Heritage Online Presenter, Computer & Graphics, Volume 52, page 129--141, 2015
- Soler, F., Melero, F. J., & Luzón, M. V. (2017). A complete 3D information system for cultural heritage documentation. Journal of Cultural Heritage, 23, 49-57.
- Stefani, C., Brunetaud, X., Janvier-Badosa, S., Beck, K., De Luca, L., Al-Mukhtar, M.: Developing a toolkit for mapping and display stone alteration on a web-based documentation platform. J. Cult. Heritage 15(1), 1–9 (2014)
- Shi, Weiqi. et al. 2016. CHER-Ob: A Tool for Shared Analysis in Cultural Heritage, proceedings of EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2016).

5.1.2.2 Visualization and analysis

- M. Callieri M., M. Dellepiane, P. Cignoni, R. Scopigno, "Processing sampled 3D data: reconstruction and visualization technologies", Chapter in "Digital Imaging for Cultural Heritage Preservation", F. Stanco, S. Battiato, G. Gallo (Ed.s), Taylor and Francis, page 103--132 - 2011.
- P. Cignoni, C. Montani, C. Rocchini, R. Scopigno “External Memory Management and Simplification of Huge Meshes”, IEEE Trans. on Visualization and Computer Graphics, vol. 9(4),Oct-Dic 2003, pp. 525-537.
- M. Doerr, M. Theodoridou, “CRMdig: A generic digital provenance model for scientific observation”, TaPP, 2011.

- D. Koller, B. Frischer, G. Humphreys. "Research challenges for digital archives of 3D cultural heritage models." journal on computing and cultural heritage (JOCCH) 2.3 (2009): 7.
- D. Koller, M. Turitzin, M. Levoy, M. Tarini, G. Croccia, P. Cignoni, R. Scopigno “Protected Interactive 3D Graphics Via Remote Rendering”, ACM Trans. on Graphics, vol. 23(3), 2004, pp. 695-703.
- A. Koutsoudis, C. Chamzas, “3D pottery shape matching using depth map images”, Journal of Cultural Heritage, Volume 12, Issue 2, April–June 2011, Pages 128-133
- Leoni C., Callieri M., Dellepiane M., O'Donnell D., Rosselli Del Turco R., Scopigno R. “The dream and the cross: a 3D Scanning project to bring 3D content in a digital edition”. In: ACM Journal on Computing and Cultural Heritage (JOCCH), vol. 8 (5) article n. 5. ACM, 2015.
- Pietroni N., Corsini M., Cignoni P., Scopigno R., “An interactive local flattening operator to support digital investigations on artwork surfaces”. IEEE Trans. on Visualization and Computer Graphics. 2011. Vol.17(12):1989-96.
- D. Pitzalis et al. "LIDO and CRM dig from a 3D cultural heritage documentation perspective." Proceedings of the 11th International conference on Virtual Reality, Archaeology and Cultural Heritage. Eurographics Association, 2010.
- F. Ponchio, M. Dellepiane, Multiresolution and fast decompression for optimal web-based rendering , Graphical Models, Volume 88, page 1-11 - November 2016.
- R. Scopigno, M. Callieri, P. Cignoni, M. Corsini, M.Dellepiane, F. Ponchio and G. Ranzuglia, "3D models for Cultural Heritage: beyond plain visualization", IEEE Computer, July 2011, vol. 44 no. 7, pp. 48-55.
- Shi, Weiqi. et al. 2016. “CHER-Ob: A Tool for Shared Analysis in Cultural Heritage,” proceedings of EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2016). http://graphics.cs.yale.edu/site/publications/cher-ob-tool-shared-analysis-cultural-heritage
- F. Uccheddu, M. Corsini, and M. Barni. "Wavelet-based blind watermarking of 3D models." Proceedings of the 2004 workshop on Multimedia and security. ACM, 2004.
- Zhang, Y., et al. "Classical sculpture analysis via shape comparison." Culture and Computing (Culture Computing), 2013 International Conference on. IEEE, 2013.

Online 3D viewers

Commercial systems

- 3D Viewer online https://www.3dvieweronline.com/
- Autodesk A360 viewer https://a360.autodesk.com/viewer/
- GrabCAD https://grabcad.com/
- P3D.in https://p3d.in
- Share my 3D https://www.sharemy3d.com/
- Sketchfab https://sketchfab.com/
- STL Viewer http://www.viewstl.com/

Academic/open sources platforms

- 3D Hop http://3dhop.net/ - Potenziani M., Callieri M., Dellepiane M., Corsini M., Ponchio F., Scopigno R., 3DHOP: 3D Heritage Online Presenter, Computer & Graphics, Volume 52, pp. 129--141, 2015
- ARIADNE's Visual Media Service http://visual.ariadne-infrastructure.eu/
- PoTree http://potree.org/demo/plyViewer/plyViewer.html
- OpenJscad http://openjscad.org/
- Smithsonian X3D (powered by Autodesk) https://3d.si.edu/
- X3DOM https://www.x3dom.org/ - J. Behr et al. 2015. webVis/instant3DHub: visual computing as a service infrastructure to deliver adaptive, secure and scalable user centric data visualisation. In Proceedings of the 20th International Conference on 3D Web Technology (Web3D '15). ACM, pp. 39-47.

Interlinking 3D objects to other media

- The Culture 3D Cloud project developed at CNRS/MAP lab : http://c3dc.fr/
- The Aioli platform developed at CNRS/MAP lab
- The CHER-Ob platform developed at Univ. Yale, Computer Graphics Group:
  - Shi, Weiqi. et al. 2016. “CHER-Ob: A Tool for Shared Analysis in Cultural Heritage,” proceedings of EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2016)
  - http://graphics.cs.yale.edu/site/cher-ob-open-source-platform-shared-analysis-cultural-heritage-research

5.1.2.3 Description and preservation

Metadata for 3D Model Long Term Preservation[98]

- CINES 2014a. Le concept d'archivage numérique pérenne. https://www.cines.fr/archivage/un-concept-des-problematiques/le-concept-darchivage-numerique-perenne/ (site accessed on 07/04/2016), last update 22/04/2014.
- Denard, Hugh "A New Introduction to the London Charter" inA.Bentkowska-Kafel, D. Baker & H. Denard (eds.) Paradata and Transparency in Virtual Heritage Digital Research in the Arts and Humanities Series (Ashgate, 2012) 57-71. http://www.londoncharter.org/ (site accessed on 07/04/2017)
- Fernie, K., Gavrilis, D., Angelis, S., 2013. The CARARE metadata schema, v. 2.0.
- Peña Serna, S., Scopigno, R., Doerr, M., Theodoridou, M., Georgis, Ch., Ponchio , F., & Stork , A. (2011). 3D-centered media linking and semantic enrichment through integrated searching, browsing, viewing and annotating. Proceedings of VAST11: The 12th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage, Prato, Italy, October 18-21, 2011.
- Ronzino, P., Hermon, S., Niccolucci, F., A metadata schema for cultural heritage documentation, V., Capellini (ed.), Electronic Imaging & the Visual Arts: EVA, 2012.

Key metadata schemas for 3D models

- ARCO (Augmented Representation of Cultural Objects) : M. Patel, M. White, K. Walczak, and P. Sayd, "Digitisation to presentation: Building virtual museum exhibitions," in Vision, Video and Graphics, 2003.
- CARARE 2.0 (3D-ICONS): http://3dicons-project.eu/eng/Resources/Documentation/CARARE-2.0-schema
- CRMDIG : http://www.ics.forth.gr/isl/index_main.php?l=e&c=656
- LIDO : http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/
- METS: http://www.loc.gov/standards/mets/
- The STARC Metadata schema : Ronzino, P., Hermon, S., Niccolucci, F., A metadata schema for cultural heritage documentation, V. Cappellini (ed.), Electronic Imaging & the Visual Arts: EVA, 2012.

[98] References compiled by Sarah Tournon-Valiente (Archeovision Lab., CNRS, France)

5.1.2.4 Bridges between Cultural and Museology

3D digitization of Nantes historic harbour[99]

- B. Guillet, C. Courtin, F. Laroche, J.-L. Kerouanton, Nantes 1900 - la maquette du port, ouvrage collectif, Musée d'histoire de Nantes, France, 88 p., 978-2-906519-49-7.
- B. Hervy, F. Laroche, A. Bernard, Framework for historical knowledge management in museology, Int. J. Product Lifecycle Management, Inderscience, 2017, vol. 10 (1), pp.44-68, DOI: http://dx.doi.org/10.1504/IJPLM.2017.083001.
- B. Hervy, F. Laroche, J.-L. Kerouanton, A. Bernard, C. Courtin, L. D'Haene, B. Guillet, A. Waels, Museum augmented interface for historical scale models: towards a new way for cultural heritage promotion, International Journal of Virtual Reality, volume 15 (1), pp.3-9, ISSN: 1081-1451.
- N. Ma, F. Laroche, B. Hervy, J.-L. Kerouanton, Virtual conservation and interaction with our cultural heritage: Framework for multi-dimension model based interface, Digital Heritage International Congress, Marseille, pp. 323 – 330, DOI : 10.1109/DigitalHeritage.2013.6743756.
- A. P. Michel , S. Kilouchi, « Renault-Billancourt' C5 Workshop in the Digital Age: a New Story of the 1922 Assembly Line », Actes du Symposium, L'histoire contemporaine à l'ère numérique, Bruxelles, PIE-Peter Lang, 2013, p. 207-221.

Repositories for 3D models and searching features[100]

- Kristian Hildebrand and Marc Alexa. 2013. Sketch-based pipeline for mass customization. In ACM SIGGRAPH 2013 Talks (SIGGRAPH '13). ACM, New York, NY, USA, , Article 37 , 1 pages. DOI=http://dx.doi.org/10.1145/2504459.2504506
- I. K. Kazmi, L. You and J. J. Zhang, "A Survey of 2D and 3D Shape Descriptors," Computer Graphics, Imaging and Visualization (CGIV), 2013 10th International Conference, Macau, 2013, pp. 1-10. doi: 10.1109/CGIV.2013.11
- Klavans, J. L., LaPlante, R. and Golbeck, J. (2014), Subject matter categorization of tags applied to digital images from art museums. J Assn Inf Sci Tec, 65: 3–12. doi:10.1002/asi.22950
- Liu, ZB., Bu, SH., Zhou, K. et al. A Survey on Partial Retrieval of 3D Shapes. J. Comput. Sci. Technol. (2013) 28: 836. doi:10.1007/s11390-013-1382-9

[99] References compiled by Florent Laroche (Ecole Centrale de Nantes, LS2N, France)

[100] References compiled by K. Rodriguez (Univ. Brighton, UK)

- David Lo Buglio, Vanessa Lardinois, and Livio De Luca. 2015. What Do Thirty-One Columns Say about a “Theoretical” Thirty-Second?. J. Comput. Cult. Herit. 8, 1, Article 6 (February 2015), 18 pages. DOI=http://dx.doi.org/10.1145/2700425
- Pedro Pascoal, Alfredo Ferreira and Jorge, Joaquim. (2012) Towards an Immersive Interface for 3D Object Retrieval. Eurographics Workshop on 3D Object Retrieval. The Eurographics Association
- Sfikas, K., Pratikakis, I., Koutsoudis, A. et al. Partial matching of 3D cultural heritage objects using panoramic views. Multimed Tools Appl (2016) 75: 3693. doi:10.1007/s11042-014-2069-0
- Karina Rodriguez Echavarria and Ran Song. 2016. Analyzing the Decorative Style of 3D Heritage Collections Based on Shape Saliency. J. Comput. Cult. Herit. 9, 4, Article 20 (December 2016), 17 pages. DOI: https://doi.org/10.1145/2943778
- Pena Serna, Sebastian and Scopigno, Roberto and Doerr, Martin and Theodoridou, Maria and Georgis, Christos and Ponchio, Federico and Stork, Andre (2011). 3D-centered Media Linking and Semantic Enrichment through Integrated Searching, Browsing, Viewing and Annotating. In VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage
- Thomas Steiner, Lorenzo Sutton, Sabine Spiller, Marinella Lazzaro, Francesco Nucci, et al. I-SEARCH - a multimodal search engine based on rich unified content description (RUCoD). WWW 2012 Companion - 21st international conference companion on World Wide Web, Apr 2012, Lyon, France. ACM, pp.291-294, 2012.

5.1.3 Ongoing efforts

5.1.3.1 Production and processing

Incorrect data acquisition processes can make an effective use of 3D models very hard or impossible. Multiple technologies are available, and none of which is the solution for all problems. Making a correct choice given a specific task is thus not an easy task for the users. More training and guidance are much needed on the following topics:

- Guidance on how the critical issue is to prepare the scene before digitization to make more evident the information we want to sample (archaeology case presented by Dell’Unto and others).
- Digitization is already an interpretation. It is thus critical to drive the digitization, ensuring that the digitization action focuses on the important areas. Note that

evaluating the importance cannot be demanded of the technologists, as the digitizing practitioner should understand the knowledge behind the sampled surface to sample it correctly.

How to ensure/enforce the above issues? How could PARTHENOS contribute to improve ability of users in making a correct and qualified use of existing technologies?

- Planning and implementing training events?
- Producing Guidelines? Maybe define an improved planning protocol, where the digitization is first planned jointly on a graphical reference (a map or an image) where the excavation people define which should be the focus of the digitization and the areas that should be given priority or more attention… Some sort of quick annotation done on the field that could drive the work of the digitization people. These annotations could also be projected back on the resulting 3D models (possible if annotation is done on registered images, following the approach proposed by MAP-CNRS). Panoramic images can play an important role in this process (since they are now easy to acquire and give a global view over the working area).
- Move digitization from the hands of the technologist, have it driven by domain experts (archaeologist, restorers)?

5.1.3.2 Visualization and analysis

The impact of the themes treated in this section for the PARTHENOS community and work programme are listed in the following. We list the sub-domains discussed and for each of them we define the possible outcomes or actions that could be activated in PARTHENOS. Those actions could be related to dissemination of knowledge (training) or to research/technological transfer efforts (development).

Web-based visualization technologies and related issues

- Training actions (remember the usual need of building a common language linking technologists and DH people; we should plan training actions at two levels: beginning/advanced)
- Production of Recommendations / How to guides, to disseminate and consolidate best practices.

- Devise proof of concepts for visualization tools managing diverse types of datasets or for collections of datasets?

3D browsers: plain Visualization vs. Analysis approaches

- Tools development: move from web tools implementing generalistic visualization functionalities, to more specific analysis tools, fulfilling the needs of the DH community
- Training actions: enabling our community to use and deploy existing web based resources

Local processing vs. CLOUD processing

- Tools development (or setting up initiatives for sharing available resources)
- Training actions

Interlinking 3D objects to other media

- Managing multimodal data with integrated tools or repositories is quite a new approach in DH. We have some tools available (even in the open source domain) for which training effort should be invested. In some other cases, we could envision some development of new tools.

Search and retrieval over DB/Archives of 3D shapes

- On this subject we need further study and research, both at the technology and humanities level: find a clear justification and research objectives according to DH user requirements, find proper use cases, considerably improve current technologies.

Encoding/including time

- On this subject we need further study, to find a clear justification according to user requirements, to verify if a common way to manage time could be shared between different stakeholders (should we develop a model customized towards the requirements of the archaeology domain, or following a set of requirements common to several sub-domains?)

5.1.3.3 Description and preservation

PARTHENOS as an inter-disciplinary network of Research Infrastructures has the possibility to take a new, wider view of the question of metadata for 3D models by building from the multi-disciplinary base it provides to set up a generic perspective on the needs for describing 3D models in their full provenanced context.

The decision of which schema or schemas to recommend falls within the remit of WP4. The work undertaken during this session and followed up by the research of relevant previous projects and existing schema should be combined with an analysis of their use/usefulness as discovered empirically by analysis against existing repositories of 3D model data within and outside of the PARTHENOS network. This list remains to be compiled and used.

Another important resource to keep in the loop with regards to the progress of this task are the WP6 implementers of PARTHENOS who could provide search and display functions based off these metadata recommendations.

On basis of the above, a first draft recommendation on schema/ strategy for 3D object metadata could be articulated. An open discussion was whether a tool for generating/managing such metadata could be within the scope of PARTHENOS.

5.1.3.4 Bridges between Cultural and Museology

The issues raised by the reuse of 3D objects relate closely to the use of appropriate standards and methods able to guarantee their long-term usability. This need has been stressed in all the previous sections. However, the participants of the session "3D, Cultural Heritage and Museology" agreed that some efforts should be put to develop best practices related to use and reuse policies of 3D objects. These best practices could take the form of a guide, helping researchers defining the access and reuse policy of the material they produced. PARTHENOS is currently developing such a guide within WP3, as an interactive Common Policies Wizard. In this broad context, 3D specialists should establish a more specific best practices guide regarding the access and the reuse of the objects they produced. Some participants in the workshop, also PARTHENOS members, have the opportunity to contribute to the building of such guidelines in close interaction with the development of the Common Policies Wizard. The solicitation of transdisciplinary research programs, which associate technological, conceptual and knowledge challenges, may be a way to consolidate and disseminate these best practices.

About the level of the dissemination of 3D objects to a wider public, it is stated that scientific 3D imaging should be presented in a specific visual language, and be inserted in public usage scenarios precisely designed. Sketching these design methods is still a pending task, which should be based on the necessity to render properly the scientific reasoning rigour and the subtles points expressed by the hypotheses. A crucial question is the responsibility of this task. Should it be carried out by the scientifics labs, considering it as a significant part of the research project, or by other (private) structures, with all the risks this entails. This question is highly decisive for Archaeology and Cultural Heritage, as they face a high demand from the society (education, tourism, local development, etc). The multifaceted nature of 3D models offers new possibilities in this domain, allowing for new ways of knowledge sharing.

Finally, the Data Reuse Charter, a service developed by PARTHENOS, amongst other partners, could be a real added value to the reuse of 3D objects. Its aim is to develop an environment to set out the conditions of collaboration between Cultural Heritage Institutions and scholars. It simplifies information retrieval and transactions related to the scholarly use of cultural heritage data. The Charter does not express constraints regarding data reuse conditions, but rather reflects the actors' policies. It does encourage good practices by offering guidelines based on recognized standards. In other words, scholars that produce 3D objects could use the Charter to declare the technical and legal requirements to abide by in order to reuse such pieces of work.

5.2 Introducing standards for the characterization of cultural heritage materials and artefacts

5.2.1 Raman microspectrometry for the analysis and identification of pigments[101]

5.2.1.1 Scope

The specific standard describes a detailed methodology to record Raman spectra of colour painted materials and artworks for the non-invasive identification of organic and inorganic pigments. This document will present standard protocols that can be applied in different

[101] Panayiotis Siozos (FORTH)

types of Raman instruments (bench-top and portable) and in various types of laser sources.

Concerning the choice of material/artefact, our suggestion, is to consider, for example, painted stone sculpture.

This method may be applied to:

- painted artefacts either untreated or subjected to any treatment or ageing
- representative coloured surfaces of objects.

5.2.1.2 Technical overview

Raman spectrometry is a widely used technique for examining various types of cultural heritage materials and artworks, with significant contributions in studies concentrating on pigment identification. The Raman effect provides a quick and relatively straightforward molecular identification of a material under examination due to the fact that the Raman spectrum can be considered as a fingerprint that is used for compound identification. The method is considered significantly versatile therefore different types of equipment are commercially available:

1) Bench-top microscopes: This type of Raman instrument is found in a laboratory and is the instrument with the highest performance in terms of speed, signal intensity, spatial and spectral resolution, stability and freedom from disruptive vibrations. A wide variety of lasers can be installed in such an instrument and the use of a microscope also ensures that a very small area is analysed each time in the range of few micrometres across and in depth. Due to this the interference of surrounding materials is limited.

2) Probe instruments: This is an easily transportable piece of equipment that can be used on site during an excavation or on unmovable objects such as wall paintings, cave paintings, mosaics etc. Probe instruments have significant compared to table top such as reduced signal intensity, spatial and spectral resolution, a limited choice in terms of lasers, a less-than-ideal ability to view and evaluate with a proper microscope the sample under examination, and the presence of vibrations that can hinder the analysis.

3) Handheld instruments: This type of instrument is easy to use and is especially suited for the analysis of inorganic materials, for example during a mineral survey in the field.

However, it has an even more limited spatial and spectral resolution compared to the probe. It may also be difficult to set the power intensity at the sample to a suitable level, and having no microscope objective at the end of the instrument, a handheld Raman probe does not allow the analyst to inspect and choose the target region carefully.

Types of excitation sources (lasers)

Most of the lasers commonly used in a cultural heritage laboratory are in the visible (blue, green, red) and near infrared red range. Laser sources in the UV are becoming more common in Raman instruments. The Raman effect is based on light scattering, for that reason absorption of the laser beam by the sample needs to be limited. The intensity of the Raman signal is inversely proportional to the excitation wavelength, thus the Raman signal generated by a blue excitation source is more intense compared to the signal produced by a red excitation source.

The available types of laser sources in the blue spectral range are:

- Blue and green lasers: 1) argon ion laser (488 nm), 2) argon ion (514.5 nm) 3) second harmonic of Nd:YAG laser, 532 nm, 4) diode lasers.

- Red lasers: 1) He/Ne laser (632 nm), 2) krypton ion laser (647 nm) or any other red solid state laser and 3) diode lasers.

- Near infrared lasers: 1) Diode lasers usually between 780 and 785 nm, 2) fundamental of Nd:YAG laser at 1064 nm (usually found on FT-Raman spectrometers).

Intensity protocol for damage protection

Significant absorption of the laser beam by the material is likely to generate local overheating, which in turn can produce chemical alteration and degrade the irradiated area of the sample under examination. The laser-degraded area may also give rise to a Raman spectrum which can be incorrectly attributed as a different material. Therefore, significant precautions must be taken into account during measurement to avoid such kind of incidents.

An important procedure must be followed during analysis of the material/artwork under Raman analysis. The measurement must begin using a very low laser power (well below 1 mW) and progressively increased as needed to obtain a good spectrum, always making sure that the sample is not being damaged by the laser irradiation. Commercial Raman instruments usually come equipped with a set of neutral density filters that allow adjusting the laser power at the sample, reducing it from 100% to below 1% of the maximum intensity. It should be noted that laser-induced overheating can occur even if there are not any visible signs of alteration on the sample. To confirm that the temperature in the sample is sufficiently low, the Raman peaks in the spectrum have to remain unchanged during measurement. In the case that the Raman bands alter during measurements, it is likely the specimen is being adversely affected by the laser beam and the measurement must stop immediately.

Fluorescence effect

The presence of fluorescence in the Raman signal can severely compromise the spectral characteristics of the studied sample/material. In many cases, fluorescence is generated from the presence of organic materials containing chromophores that can be excited at different wavelengths in the visible spectral range. These substances can themselves be responsible for the colour or can be used as binding media mixed with inorganic pigments. However, a fluorescent background can frequently be observed also for inorganic materials. Currently, there is not a direct way to overcome the fluorescence emission in the Raman spectra. However, fluorescence can be reduced by using a laser source in the longer wavelength (near infrared). In this case the Raman signal is also reduced, therefore a specific wavelength has to be selected in order to achieve highest Raman signal with low fluorescence background.

Analysis of spectra: Identification of compounds

Identifying compounds on the basis of their Raman bands can be a complex operation, which requires a detailed knowledge of group theory and involves lengthy calculation. The identification of compounds using theoretical calculations is rarely performed and the most frequent procedure followed is based on the comparison of the wavenumbers of the Raman peaks with the wavenumbers from Raman spectra in existing databases and literature resources. Most of these resources are available in published domains and can be used as a reference (see resources section). During this procedure a graph is created that contains the Raman spectrum from the studied material and one or more Raman spectra from the reference materials. The wavenumbers of the Raman peaks are also presented in the graph and explained in the graph’s caption.

5.2.1.3 Resources

- Analytical Methods Committee, AMCTB No 67, Raman spectroscopy in cultural heritage: Background paper, Anal. Methods, 2015,7, 4844-4847. DOI: 10.1039/c5ay90036k

- Databases and resources of Raman Spectra

Material	**Reference**
Pigments, minerals	I. M. Bell, J. H. Clark, P. J. Gibbs, “Raman spectroscopic library of natural and synthetic pigments (pre-~1850 AD)”, Spectrochim. Acta A **53**, 2159-2179 (1997). Also at: http://www.chem.ucl.ac.uk/resources/raman/index.html
Pigments, minerals, pigment media and varnishes	L. Burgio, R. J. H. Clark, “Library of FT-Raman spectra of pigments, minerals, pigment media and varnishes, and supplement to existing library of Raman spectra of pigments with visible excitation”, Spectrochim. Acta A **53**, 1491-1521 (2001).
Natural organic binding media and varnishes	P. Vandenabeele, B. Wehling, L. Moens, H. Edwards, M. DeReu, G. Van Hoydonk, “Analysis with micro-Raman spectroscopy of natural organic binding media and varnishes used in art”, Analytica Chimica Acta **407**, 261-274 (2000).
Modern azo pigments	P. Vandenabeele, L. Moens, H. G. M. Edwards, R. Dams, “Raman spectroscopic database of azo pigments and application to modern art studies”, J. Raman Spectrosc. **31**, 509-517 (2000).
Colored glazes	P. Colomban, G. Sagon, X. Faurel “Differentiation of antique ceramics from the Raman spectra,of their colored glazes and paintings” J. Raman Spectrosc. **32**, 351-360 (2001)

Material	**Reference**
Minerals, metal corrosion products	M. Bouchard, D. C. Smith, “Catalogue of 45 Raman spectra of minerals concerning research in art history or archaeology, especially on corroded metals and coloured glass”, Spectrochim. Acta. A **59**, 2247-2266, (2003)
Minerals	California Institute of Technology, Division of Geological and Planetary Sciences (USA) (http://minerals.gps.caltech.edu/files/raman/) The RRUFF project, Univ. of Arizona (USA) (http://rruff.info/)
Artists’ and related materials	The Infrared and Raman Users Group (IRUG) Spectral Database (http://www.irug.org)
Artists’ and related materials	e-VISART Database, Univ. of the Basque Country, Dept. of Analytical Chemistry (Spain) (http://www.ehu.es/udps/database/database.html)
Organic compounds	Spectral Database for Organic Compounds, AIST (Japan) (http://riodb01.ibase.aist.go.jp/sdbs/)

Table 4 - Databases and resources of Raman Spectra

5.2.1.4 Ongoing efforts

The list of available databases of Raman spectra and multispectral imaging sources will be extended. Furthermore, a document describing in detail the standard procedure that has to be followed for reliable Raman characterization of artworks and cultural heritage materials will be produced and included in the “Introducing standards for the characterization of cultural heritage materials and artefacts” document (see Appendices). The procedure that has been already prepared for “Multispectral imaging measurements of painted stone sculpture” will be expanded. Finally, a significant effort will be applied in order to increase the available cases for preparation standard documents for the characterization of materials and artworks, particularly preparing standards related with the application XRF analysis in the field cultural heritage.

5.2.2 Multispectral imaging for surface mapping of pigments[102]

5.2.2.1 Scope

This standard describes a method to record multispectral images of colour painted materials and artworks, which is a commonly used technique currently available to the scientist, conservator, archaeologist and art historian for the non-invasive investigation of works of art. This document will concentrate on the wavelength range that can be observed using modified commercially available cameras, which typically employ silicon based sensors sensitive from approximately 350 nm to 1100 nm. Cameras based on InGaAs sensors, which can record infrared radiation from approximately 700 nm to 1700 nm, can be used regularly in cultural heritage applications but due to their specialized technology they are out of the scope of this standard.

Concerning the choice of material/artefact, our suggestion, is to consider, for example, painted stone sculpture.

This method may be applied to:

- painted artefacts either untreated or subjected to any treatment or ageing
- representative surfaces of objects, indoors or outdoors.

5.2.2.2 Technical overview

Multispectral imaging is the procedure used to observe an object using selected ranges of wavelengths in the electromagnetic spectrum that include and extend beyond the capabilities of the human eye. A generic setup for multispectral imaging is composed of three main components:

1) Incoming radiation, which is generated by a radiation source and travels towards the object;
2) The object, which interacts with the incoming radiation;
3) Outgoing radiation, which, following the interaction between the incoming radiation and the object, travels from the object to the recording device.

The extent to which this radiation will penetrate the object under investigation will be dependent on its wavelength and on the absorbance of the materials which compose the object, with longer wavelengths of radiation generally penetrating further into the piece. For example, when examining a painting, shorter wavelengths (such as UV) are often

[102] Mikel Sanz (CSIC)

readily absorbed by the outer layers (usually varnishes), while longer wavelengths can pass through the varnish and interact with the pictorial film and the under drawing. The radiation reaching any particular point in the object can be: (i) absorbed, (ii) reflected and/or (iii) absorbed and re-emitted as luminescence at longer wavelengths. Each outcome produces an image set which yields information specific to that point. Thus by selecting particular combinations of illumination and detection ranges, it is possible to gain insight about the distribution of materials in the object under study.

The test equipment for the acquisition of multispectral images has to be made up of a number of equipment components: 1) Radiation sources (RS) which provide the incident radiation to the object being studied, 2) a filter or set of filters in order to allow the transmittance of radiation in the wavelength range under study (FT) and exclude unwanted radiation from being recorded (FR); 3) a detector or recording device, a commercially available digital camera (modified by removal of the IR-blocking filter in the case of IR-induced or IR-reflected images and 4) a set of standards to enable the post-processing methods. The possible image set with the recommended equipment components are:

- Visible-reflected (VIS) image corresponds to standard photography and records the reflected light in the visible region (400-700 nm) from an object when this is illuminated with visible light. The image is collected in the range in which the object is usually observed and can serve as the reference point to interpret the other image sets. RS: Tungsten Halogen or Xenon lamps or LED source, FT: None, FR: Bandpass filters to allow only light in the range 400-700 nm.

- Infrared-reflected (IRR) images record the reflected radiation in the infrared region (700-1100 nm) from an object when this is illuminated with infrared radiation. This image can be valuable in revealing under drawings and concealed features. RS: Tungsten Halogen lamp, FT: None, FR: Bandpass filter to allow only light in the range 700-1100 nm.

- Ultraviolet-reflected (UVR) image records the reflected radiation in the ultraviolet region (200-400 nm) from an object when this is illuminated with ultraviolet radiation. RS: Black light fluorescent lamp or UV LED source, FT: Shortpass infrared filter (<700 nm), FR: Bandpass filters to allow only light in the range 200-400 nm.

- Ultraviolet-induced luminescence (UVL) image records the emission of light (luminescence) in the visible region (400-700 nm) from an object when this is illuminated with UV radiation. This image is used to investigate the distribution of luminescent materials, such as organic binders and colourants. RS: Black light fluorescent lamp or UV LED source, FT: None, FR: Bandpass filters to allow only light in the range 400-700 nm.

- Visible-induced infrared luminescence (VIL) image records the emission of radiation (luminescence) in the infrared region (700-1100 nm) from an object when this is illuminated with visible light. RS: Incandescent lamp or Visible LED source, FT: Infrared shortpass filter (<700nm), FR: Longpass filter (>700 nm) to allow only infrared light (700-1100 nm).

- Visible-induced visible luminescence (VIVL) image records the emission of light in the visible region (500-700 nm) from an object when this is illuminated with visible light (400-500 nm). RS: Blue LED source, FT: None, FR: Bandpass filter (500-700 nm)

In all the cases, Standards and Calibration targets must be used. They are uniform reflective boards that should be a grey Lambertian reflector: a surface showing the same radiance when viewed from any angle. Several commercial available products are available :

1) Spectralon diffuse reflectance standards

- http://www.labsphere.com/labsphere-products-solutions/materials-coatings-2/targets-standards/diffuse-reflectance-standards/diffuse-reflectance-standards/
- http://www.labsphere.com/products/reflectance-standards-and-targets/reflectancetargets/spectralon-targets.aspx

2) Macbeth (X-rite) ColorChecker Chart

- http://xritephoto.com/ph_product_overview.aspx?catid=28
- http://xritephoto.com/ph_product_overview.aspx?id=1257&catid=28

3) Uniform reflective board

- http://xritephoto.com/ph_product_overview.aspx?id=944&catid=28

- http://www.labsphere.com/labsphere-products-solutions/materials-coatings-2/targets-standards/diffuse-reflectance-standards/color-standards/
- http://www.labsphere.com/products/reflectance-standards-and-targets/reflectancetargets/spectralon-targets.aspx

5.2.2.3 Resources

For a general discussion and recommendations on the safe handling and positioning of objects for imaging in art historical and conservation contexts, the reader is referred to the AIC Guide to Digital Photography and Conservation Documentation
http://cool.conservation-us.org/coolaic/sg/emg/dtf/DTF_Online_Weblinks.pdf
For a free image processing system that includes a range of filters, arithmetic operations, colour processing, histograms, and geometric transforms
http://www.vips.ecs.soton.ac.uk/index.php?title=VIPS

- Data repository

http://www.labsphere.com/support/datasheets-library/

- Bibliography

https://www.britishmuseum.org/pdf/charisma-multispectral-imaging-manual-2013.pdf

- Blog entries

For information about new features in nip2 and how to report any problems encountered with nip2: http://libvips.blogspot.com.es/?view=magazine