This project is funded under the
ICT Policy Support Programme part of the
Competitiveness and Innovation Framework Programme.

Project

Project Acronym:	AthenaPlus
Grant Agreement number:	325098
Project Title:	Access to cultural heritage networks for Europeana

Deliverable

Deliverable name:	**First release GLAM sector reference terminologies**
Deliverable number:	**D4.1**
Delivery date:	September 2013
Dissemination level:	Public
Status	Final
Authors (organisation)	Eva Coudyzer (KMKG-MRAH)
Contributors (organisation)	
Reviewers (organisation)	Bert Lemmens (PACKED), Marie-Véronique Leroi (MCC)

Revision History

Revision	Date	Author	Organisation	Description
V0.1	2013-09	Eva Coudyzer	KMKG	Draft version
V0.2	2013-09	Marie-Véronique Leroi	MCC	Suggestions
V0.3	2013-09	Bert Lemmens	PACKED	Reviewer: suggestions and approval
V1	2013-09	Maria Teresa Natale	ICCU	Formal check

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Project Coordinator: Istituto centrale per il catalogo unico delle biblioteche italiane
Address : Viale Castro Pretorio 105 – 00185 Roma
Phone number : +3906 06 49210 425
E-mail : info@athenaplus.eu
Project WEB site address : http://www.athenaplus.eu

Table of Contents

1 EXECUTIVE SUMMARY

The DoW of the AthenaPlus project describes deliverable D4.1 as follows: “a first release of GLAM-sector terminologies with a terminology resource report. It describes the results of the analyses of GLAM terminologies, the selection criteria used for the collection of suitable reference terminologies, as well as a detailed overview of the terminologies gathered in the registry of terminologies.”

The basic goal of this deliverable is to give an overview of suitable terminologies used by cultural institutions collaborating in the AthenaPlus and Linked Heritage projects. Suitable terminologies will be imported in the repository of the Terminology Management Platform (TMP), where they will be mapped to other terminologies using SKOS properties and exported in RDF so they can be reused as linked open data.

2 INTRODUCTION

The deliverable is structured as follows:

- Introduction with references to terminology practice, the semantic web (SKOS/RDF), the Linked Heritage project and a state of the art (terminology surveys, terminologies published in SKOS and interlinked thesauri or meta-thesauri).
- Results of the terminology survey conducted in July and August 2013 with information on organization, terminology type, management, standards and export formats, target groups, accessibility, rights and semantic enrichment. Some specific selection criteria are highlighted.
- Suggested strategy for importation in TMP and possible future developments
- Conclusion

2.1 Background

The deliverable D4.1 describes the resources needed for importation, mapping en export of terminologies in the Terminology Management Platform. The development of this open-source tool started in the Linked Heritage project. A production version will be accessible in month 9 of the AthenaPlus project (November 2013).

2.3 Role of this Deliverable in the Project

The deliverable is important because it analyses in detail the different terminologies used by the AthenaPlus partners. Before a terminology can be imported in the TMP, some technical and content-related criteria must be met, e.g. terms must be identified by an identifier in the CMS, the rights on the terminology must be cleared and the terminology must be a controlled vocabulary. The information in this deliverable can also be relevant to the experiments in WP4 with Linked Open Data and the re-use of cultural resources in WP5.
The conclusion of this deliverable results in an overview and planning for the work in the TMP. It will ensure a fluent workflow for the tasks in WP4.

3 Introduction

3.1 Terminology practice

In this deliverable a "terminology" is a "list of terms". This list can be simple and unstructured or "flat", but it can also contain relations between two or more terms. When the terms in a term list are structured according to hierarchical, equivalent and associative relations they are called *thesauri*, e.g. used for describing museum collections. When terms are structured following fixed numerical or lexical codes they can be *classifications*, e.g. for classifying books in libraries.
A controlled vocabulary is a structured list of descriptors, where each descriptor is a preferred term with an unambiguous, non-redundant definition. Controlled vocabularies are created following specific construction rules, established by international standard organizations such ISO and ANSI.
The principle of thesaurus management is referred to in literature as "vocabulary control". This is defined by the ANSI/NISO standard as follows: "Vocabulary control is used to improve the effectiveness of information storage and retrieval systems, Web navigation systems, and other environments that seek to both identify and locate desired content via some sort of description using language. The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval."[1]

Terminologies are developed and used by cultural organisations to support information-seeking activities. Some examples of their use are listed below:

- An *indexing* or *inventory* tool, where a set of metadata (data on data, e.g. "title") and data (e.g. "La Joconde") are collected and categorized in standardized lists
- A *searching tool* or *query information support feature* which allows users to browse complete collections exported from a local database by means of "descriptors" or "keywords"
- A facilitator for combining multiple databases or unified access to multiple databases by mapping the users' query terms to the descriptors used in each of the databases[2].
- An *interactive term suggestion feature*, where users are presented with a list of terms to choose from, being guided by the thesaurus which provides synonyms or semantically related terms.

The last decade terminologies have started playing a more prominent role in information systems, such as in the development of the semantic web, evolving more and more from an "input" to a "retrieval" tool. In the semantic web terms are referred to as *concepts* or *units of thought*, because each term (represented by a web link) contains a mass of interlinked information, including e.g. translations.

3.2 The semantic web and SKOS/RDF

The semantic web is built on the principle of sharing and reusing data on the web to achieve better search results, irrespective of language. This can be done by automatically linking "separate" data on the web through the use of web links or URIs (*see infra*). When each concept is linked as an equivalent, synonym, broader, narrower or other relation, the web can optimize search results, which will engage larger visibility and easier access to information.

Because controlled vocabularies contain a lot of information – e.g. representing a *Berline* as a sort of four-wheeled closed carriage – they are a useful tool to "spread knowledge" on the web. Researcher Ali Shiri sums up the different reasons why online thesauri are important on the web:

- The colossal growth of information resources, demanding better subject identification
- The migration of traditional information resources to the web, calling for more consistent subject approaches

[1] ANSI/NISO Z39.19-2005
[2] Soergel D. (2003), p. 6

- An urgent need for resource description and discovery through the reuse of existing information management tools such as controlled vocabularies
- Problems associated with the quality of unstructured information retrieved from the web
- The need to provide users with knowledge structures such as thesauri for rapid and easy access to better-organized information[3]

Controlled vocabularies have been identified as *knowledge organization systems* (KOS), a term coined by the Networked Knowledge Organization Systems Working Group (NKOS). The use of knowledge organization systems on the web can also be interpreted as follows: "KOS are used to organize materials for the purpose of retrieval and to manage a collection. A KOS serves as a bridge between the user's information need and the material in the collection. With it, the user should be able to identify an object of interest without prior knowledge of its existence. Whether through browsing or direct searching, whether through themes on a web page or a site search engine, the KOS guides the user through a discovery process"[4].

The World Wide Web Consortium (W3C) developed a format in RDF/XML which facilitates the encoding of information for sharing and interoperability across various computer applications. SKOS or *Simple Knowledge Organization System* was developed to "build a bridge between the world of knowledge organization systems – including thesauri, classifications, subject headings, taxonomies, and folksonomies – and the linked data community, with the goal of bringing benefits to both."[5]

SKOS provides new opportunities in the semantic web environment:

- Link several different thesauri
- Expand search functionalities through federated searching of multiple controlled vocabularies and linked data sources
- Allow for the integration of thesauri into many web-based search engines and services
- Provide semantically rich visualisation of thesauri and links between and among thesauri
- Facilitate multilingual information access and retrieval
- Provide easy access to thesauri for indexing and information representation purposes[6]

3.3 Linked Heritage (April 2011- September 2013): TMP

The Linked Heritage project investigated in 2011 the various steps necessary for making data part of the semantic web[7]. The conclusion of the survey was that an important drawback for organisations in the GLAM-sector for linking and publishing data on the web, was the little (technical) expertise in mapping vocabularies they have, as well as the lack of an accessible tool to map and export their vocabularies in the desired format.

It was decided that an open-source tool would be developed in the Linked Heritage-project: the Terminology Management Platform or TMP. The TMP has accessible functionalities for mapping and exporting data:

- Uploading terminologies from CSV and XML-files or creating terminologies from scratch
- Editing terminologies in a collaborative online-environment
- Mapping different terminologies with a drag-and-drop feature using SKOS-mapping properties
- Exporting terminologies in SKOS/RDF

A prototype of the TMP was created in the Linked Heritage-project and a production version (accessible as a web service) will be made available in month 9 of the AthenaPlus project (November 2013). At this

[3] Shiri A. (2012), p. 290
[4] Hodge G., p. 3
[5] Shiri A. (2012), p. 130-131
[6] Id., pp. 290
[7] M.-V. Leroi et al.: http://www.linkedheritage.eu/index.php?en/181/publications.

moment the TMP is in a testing phase. Partners can test the TMP and send their feedback and suggestions to Marie-Véronique Leroi (*Ministère de la Culture et de la Communication*) and Florent André (*Université de Savoie*), who are developing the tool.
The TMP can be consulted on: www.culture-terminology.org

4 State of the Art

(a) Terminology surveys
(b) Terminologies published in SKOS
(c) Interlinked thesauri or meta-thesauri

4.1 Terminology surveys

Several surveys on terminologies and terminology practice have been conducted in the past, some of which are listed below:

- EU MINERVA project, *Multilingual websites and multilingual thesauri*, 2004-2005[8]: The aim of this survey is mapping multilingual access to the European digital cultural content. Instead of creating a brand new multilingual thesaurus, it was decided to make a survey for collecting information on multilingual websites and thesauri in use. The site gives an overview of existing multilingual thesauri and webpages across Europe.
- EU Athena project, *Identification of existing terminology resources in museums*, 2009[9] investigated by means of a questionnaire de state of the art concerning terminology use in European cultural institutions. 44 terminologies were assembled (29 thesauri, 11 classification systems, 1 glossary and 3 flat lists). 66% of the terminologies were monolingual, 34 % multilingual.
- *Inventarisatie Terminologiebronnen*, a survey done by DEN in 2010[10], collected and described terminologies from The Netherlands and Belgium (Flanders) and described. The survey listed ca. 60 terminologies used in the GLAM-sector.

4.2 Terminologies published in SKOS

There are now a range of terminologies that have been published in SKOS, even though not necessarily in the GLAM-sector, e.g. there seem to be more *skossified* classifications, authority lists or subject headings in the library sector then there are in the museum sector. This is probably due to the fact that these vocabularies have since long been standardized and internationally used. Neither does publication in SKOS mean that these terminologies are *linked to other sources*, which is the basic principle of the semantic web. Some of them have been partly linked, such as the Thesaurus of Economics linked to DBPedia.

Some sites give overviews of linked open vocabularies, e.g. the W3C-webpage contains a list of terminologies from the library domain published in SKOS[11], others include the Mondeca Linked Open Vocabularies (LOV)[12] and W3C Linking Open Data Community Project[13].

Below a summary of vocabularies published in SKOS:

- AGROVOC, an agricultural thesaurus developed and used by the *Food and Agricultural Organization of the United Nations:* http://aims.fao.org/standards/agrovoc/about
- Eurovoc, the multilingual thesaurus of the European Union: http://eurovoc.europa.eu/
- *National Agricultural Library* (NAL) thesaurus, United States Department of Agriculture: http://www.nal.usda.gov/

[8] *Multilingual websites and multilingual thesauri*, Minerva survey 2004-2005, http://www.mek.oszk.hu/minerva/survey/
[9] M-V Leroi et al., *Identification of existing terminology resources in museums*, AthenaPlus 2009
[10] *Inventarisatie Terminologiebronnen*, a survey done by DEN (Digitaal Erfgoed Nederland) in 2010, http://www.den.nl/terminologiebronnen
[11] See also the W3C- Library Linked Data Incubated Group-report for more SKOS published datasets and use cases in the library domain: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/
[12] http://lov.okfn.org/dataset/lov/
[13] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

- General Environmental Thesaurus (GEMET), *European Environment Information and Observation Network:* http://www.eionet.europa.eu/gemet/about
- Thesaurus for Economics (STW), *Leibniz Information Centre for Economics:* http://zbw.eu/stw/versions/latest/about
- DDC Dewey Decimal Classification: http://www.oclc.org/dewey/resources/summaries.en.html
- National Széchényi Library's vocabularies, Hungarian National Library: http://nektar1.oszk.hu/librivision_hun.html
- DNB authority files, GND (*Gemeinsame NormDatei*), German National Library
- UNESCO-thesaurus, United Nations Educational, Scientific and Cultural Organization: http://databases.unesco.org/thesaurus/
- BnF subject headings (RAMEAU), Bibliothèque Nationale de France: http://data.bnf.fr/semanticweb-en
- Zenon archaeological thesaurus, *German Archaeological Institute*, DARIAH-De project
- Thesaurus PICO 4.3, *Portale della Cultura Italiana On-line*, Cultura Italia: http://www.culturaitalia.it/pico/thesaurus/4.3/thesaurus_4.3.0.skos.xml
- PACTOLS thesaurus, *Peuples, Anthroponymes, Chronologie, Toponymes, Oeuvres, Lieux et Sujets*, CNRS-Frantique (*Fédération de bibliothèques et de centres de documentation en archéologie*) : http://frantiq.mom.fr/thesaurus-pactols
- British Museum thesaurus, British Museum Semantic Web Collection Online: http://collection.britishmuseum.org/
- Amsterdam Museum thesaurus and person authority files, Amsterdam Museum: http://amsterdammuseum.nl/collectie and http://pro.europeana.eu/thoughtlab/linked-open-data
- VIAF, Virtual International Authority File, hosted by OCLC: http://viaf.org/
- Library of Congress Subject Headings, Library of Congress: http://www.loc.gov/aba/cataloging/subject/
- UK Archival Thesaurus (UKAT), a subject thesaurus which has been created for the archive sector in the United Kingdom: http://www.ukat.org.uk/
- Common Thesaurus Audiovisual Archives, The Netherlands Institute for Sound and Vision, Dutch archives for public broadcast television: http://datahub.io/nl/dataset/gemeenschappelijke-thesaurus-audiovisuele-archieven
- MARC Code lists (relators, geographic areas etc.), Library of Congress: http://id.loc.gov/vocabulary/relators.html
- The Dutch version of the Art and Architecture Thesaurus (RKD). It was announced that the Getty vocabularies (Art and Architecture Thesaurus, Thesaurus Geographic Names, Union List of Artist Names) will be published in SKOS and SKOS-XL formats under the ODC-BY 1.0 license: http://www.getty.edu/research/tools/vocabularies/
- Iconclass, a classification system designed for art and iconography , RKD, The Hague: http://www.iconclass.nl/home
- SENESCHAL, cultural heritage thesauri and vocabularies from English Heritage, Royal Commission on Ancient & Historical Monuments of Scotland (RCAHM) and Royal Commission on Ancient & Historical Monuments of Wales (RCAHMW): http://www.heritagedata.org/blog/

4.3 Interlinked thesauri or meta-thesauri

The key to semantic interoperability is mapping various (monolingual and multilingual) terminologies and linking different web sources. When Berners-Lee and his colleagues described in 2001 the expected evolution of the existing web to the semantic web, they said in 2006 that: “This simple idea…remains largely unrealized”. [14]

[14] Wikipedia: http://en.wikipedia.org/wiki/Semantic_Web

Researcher J. Hunter uses the term “semantic interoperability” to refer to the process of creating inter-thesaurus semantic relationships.[15] A “meta-thesaurus” is regarded by A. Shiri as a *synthesis* of existing controlled vocabularies, achieved by linking, merging and integrating these vocabularies. The constituent thesauri are mapped onto one another, creating pointers from every concept in one thesaurus to the most equivalent concept in the others. [16]

When investigating the existence of (published) interlinked semantic relations or “meta-thesauri” we conclude that there are not many examples of mapped controlled vocabularies, which would allow e.g. easy cross-browsing of various online collections.
Some examples of mapped controlled vocabularies or mapping experiments are discussed in short below: the AGROVOC-thesaurus mapping, the Amsterdam Museum Linked Open Data-project, the VMF-mapping experiment and the manual mappings in xTree (Linked Heritage project). Another mapping experiment was done in the STELLAR-project[17] in the United Kingdom.

4.3.1 The Agrovoc-thesaurus mapping[18]

The Food and Agriculture Organization of the United Nations (FAO) has experimented with mapping several thesauri using known lexical string matching techniques. The AGROVOC-thesaurus was mapped to Eurovoc, NALT, GEMET, STW, LCSH and RAMEAU (see list above for abbreviations), all of which are published in SKOS. The aim was to find exact matches as defined by SKOS-scheme using a method that can do it automatically (since manual mapping is a very time-consuming activity). In doing so all documents and resources of these organizations would be interlinked; If, e.g. a term in the AGROVOC thesaurus is linked with a term in the GEMET-thesaurus, all documents indexed by the same term in the document repositories related to AGROVOC and GEMET are also potentially linked.

The method consisted of a combination of candidate matches automatically identified and then then manually assessed, and looked at aligning techniques based on string similarity. Problems would be alignment differences in formats, structure, semantics, and concept labels with different languages and minor differences in spelling.

In short the experiment was conducted as follows: the thesauri were loaded in a local triple store (Sesame), where concepts were mapped in pairs (one concept in AGROVOC to one concept in another thesaurus), using one preferred label per concept in one common language. Between those labels string similarity measures were applied. When candidate links were found, the data were exported to a format which allows manual evaluation by an expert. When labels are not considered identical the similarity measures were applied, and the average of their value computed. For example, “Animal products” from AGROVOC and “animal product” from EUROVOC only differ by one letter, so they score high enough to pass the threshold, and can be considered as exact match.
For evaluation by an expert, following guidelines were developed:

1. Check if there are non-preferred terms (alternative labels in SKOS terminology) associated with the candidate exact match term in order to clarify the meaning. If this is not the case, then
2. Compare the matching term with other languages in common between the two thesauri, if available. AGROVOC and NALT, for example, have in common Spanish and English
3. Take a look at the concept hierarchy, i.e. mainly parent concepts, and
4. Examine definitions or scope notes of mapped concepts, if available, to verify the correctness of exact matches

The experiment resulted in a high amount (precision rate over 90%) of exact matches. It was concluded that this number was due to the fact that thesauri use standard terminology, i.e. the preferred terms are

[15] Hunter J. (2001), p. 234-253.
[16] Shiri A. (2012), p. 220-221
[17] K. May et al. (2011)
[18] A. Morshed et al. (2011)

often identical in various thesauri. Differences across thesauri are mainly due to the use of singular and plural. The few incorrect candidate matches were classified as follows:

a) Incomplete homonymy, e.g. *flavouring* in AGROVOC (refers to action of adding flour to a substance) and *flavouring* in EUROVOC (refers to the substance added)
b) Near-homonymy, e.g. *calice* (AGROVOC) and *calices* (RAMEAU): the one refers to a concept in botanical domain, the other in object liturgies.
c) False friends: similar terms, but with different meaning, e.g. *health care – health card*
d) Other cases

4.3.2 Amsterdam Museum Linked Open Data[19]

The Amsterdam Museum Linked Open Data-project published a dataset of more than 70.0000 object descriptions as LOD, including a thesaurus and authority file. The data were mapped to EDM, using DC, SKOS and RDA Group 2 elements. Vocabulary concepts were mapped to AAT, ULAN, GeoNames and DBPedia.

The Amsterdam Museum Data consist of 3 parts:

- An object metadataset consisting of metadata records for ca. 70.000 objects
- A thesaurus with 28.000 concepts used in the metadata records
- A person authority file with ca. 65.000 names

The technical process can be summarized as follows:

- The metadata, thesaurus and person authority file were harvested through OAI-PMH in XML
- XML was converted into RDF (using XMLRDF-tool)
- Resources (persons, concepts, objects etc.) were assigned URIs
- Implicit links (e.g. between objects and thesaurus concepts) were made explicit using RDF-relations.
- Literal values that represent references are replaced by the resources. Language information of the literals was added to the data in the form of language-types literals

To make the Amsterdam Museum Linked Data interoperable with EDM, two steps were taken:

- Museum objects are represented as proxy-aggregation pairs, this means that one metadata-record has one RDF-proxy-resource ánd one RDF-aggregation resource. The proxy-resource was used for object metadata (creator, dimensions etc.) and the aggregation resource was used for provenance (data provider, rights etc.) Interoperability is thus achieved without discarding the complexity of the original data.
- Mapping Amsterdam Museum specific classes and properties to those of the EDM, using *rdfs:subPropertyOf* or *rdfs:subClassOf-relations* through a schema file. Through these mappings, interoperability of the museum-specific metadata with the EDM is achieved.

The thesaurus of the Amsterdam Museum consists of 28.000 concepts represented in SKOS. These include geographic names, motifs, events etc. Most term-based thesauri have a structure based on ISO 2788-norms. The relations between terms were assigned as *skos:broader* and *skos:narrower*. In total the thesaurus consists of 601,849 RDF-triples.

The person authority file consists of 66,966 instances of *am:Persons* (a subclass of *edm:Agent*). The persons in the file are creators, annotators, owners etc. In this case 21 distinct Amsterdam Museum predicates were used: birth dates, death dates, nationality, name spellings etc. These properties were mapped to RDA Group 2-elements using 20 *rdfs:subPropertyrelations*. In total there are 301,143 RDF triples in the Person data set.

[19] V. De Boer et al. (s.d.)

Internal links were made between thesaurus concepts and objects: 558,161 links between the 73,447 proxies to thesaurus concepts. There are also 80,432 links between proxies and persons and between proxies and other proxies.
External links were achieved by using the *Amalgame alignment platform (AMsterdam ALignment GenerAtion Metatool)*[20], using string matching techniques. This required splitting the concepts into geographic and non-geographic concepts. The non-geographic concepts were mapped to the Art and Architecture Thesaurus, using a string matching algorithm. Manual evaluation revealed that 80% of matches were correct. A similar workflow was applied to the concepts mapped to GeoNames. The person authority file was mapped to Union List of Artist Names (using a slightly different approach) and to DBPedia.

4.3.3 The Vocabulary Mapping Framework[21]

The Vocabulary Mapping Framework is an open-source tool which is used as a single ontology using RDF-triples. It can be used to automatically compute the "best fit" mapping between terms in a controlled vocabulary in different metadata schemes.
The tool was tested with a selection of controlled vocabularies: CIDOC-CRM, CRM, DCMI, DDEX, FRAD, FRBR, IDF, LOM (IEE), MARC 21, MPEG21, RDD, ONIX and RDA and RDA-ONIX framework.
The initial scope of the matched vocabularies was:

Resource categories (e.g. CD, Ebook, Photograph)
- Resource-to-Resource relators (e.g. IsVersionOf)
- Resource-to-Party-Relators (e.g. Author)
- Party-to-Party relators (e.g. AffiliatedTo)
- Party categories

The vocabularies are not very homogeneous, so groups of terms with quite divers semantics from a variety of different schemes were added to the matrix to test the methodology. The matrix is a hierarchical class ontology of concepts grouped methodically using an event-based data model. The ontology can be extended as needed to provide a mapping point for any term in a vocabulary.

Basically, the tool can be used as follows: terms from vocabularies are mapped into the matrix, not mapped directly to one another. Once a term is mapped onto the matrix, the internal links of the matrix establish computable relationships with every other mapped term in the matrix. The matrix therefore presents the sum total of all mapped concepts, plus other semantic relationships between them. The matrix can then be queried, using SPARQL or another suitable language, to find the "best fit" direct mappings from one vocabulary to another.

The VMF process could be simplified as follows:

- Creating the matrix
- Mapping to VMF
- Mapping scheme to scheme

4.3.4 Mapping experiment in xTree

xTree is a mapping tool created by the German digiCULT Museen-project, sponsored by the European Fund for Regional Development (EFRE). In the Linked Heritage-project, the tool was used to experiment with mapping terminologies. A few terminologies were imported or created in xTree: the vocabulary of the Israel Museum, the British-Museum thesaurus, the thesauri of the KMKG-MRAH, Stedelijke Musea Mechelen and Plantin-Moretus Prentenkabinet.

In contrast to the projects summarized before, the mapping in Xtree using SKOS-properties was done *manually*, which requires great investments in time and people. The thesauri of Israel Museum has been mapped to the thesaurus of the British Museum. The KMKG-thesaurus was mapped to the British

[20] semanticweb.cs.vu.nl/amalgame/
[21] The Vocabulary Mapping Framework (VMF): an introduction, v.1.0 (2009)

Museum thesaurus. The British Museum Thesaurus was used as a type of ‘reference’-thesaurus where other, smaller thesauri could be mapped to.

In a next phase an export in SKOS/RDF should be done of these vocabularies (including the mapping elements), where they will be imported in the TMP and mapped to more sources.

5 Terminology Questionnaire

5.1 Selection criteria

Before discussing the questionnaire itself, it is important to focus on some selection criteria which deserve special attention when setting up a registry of controlled vocabularies from partners who wish to import and map their vocabularies in the TMP.

5.1.1 Is the terminology a controlled vocabulary?

The more information a vocabulary contains, the more valuable it is to the semantic web. This is why we need to look at the way the information is organized in the terminology or KOS: what kind of relations does it represent? Are there definitions or scope notes? How many terms does the terminology contain? Information is stored in thesauri in a system of equivalent (preferred/alternative terms) hierarchical (broader/narrower terms) and associative (related term) relations. Additional information is stored in scope notes, history notes or editorial notes.

The ANSI/NISO standard describes the advantages and disadvantages of thesauri as follows:

- Good information about hierarchical relationships among terms
- Good information about relationships among terms
- Entry vocabulary to help users locate the correct terms
- Thesauri are useful for both indexers and searchers who need to discover the most appropriate, specific terms for their purposes
- Thesauri are time-consuming and labour-intensive to develop and maintain[22]

Considering this, a thesaurus is to be given preference to a flat term list, which contains no relations at all.

Researcher U. Miller says that thesauri should be constructed on the basis of the maximum possible number of terms and their synonyms, object relations between terms, multiple languages, and receptivity to new terms [23]. Search behavior studies have also shown that equivalent relations (or *synonyms, quasi-synonyms, abbreviations*) in a controlled vocabulary are very important. Morville and Rosenfeld explain that the mapping of many synonyms or word variants onto one preferred term or concept is an important feature allowing users to deal with the ambiguities of language during their searching and finding experience.[24]

Another interesting research project showed that the terminology used by humanities researchers was remarkably different from the vocabulary used in other fields, as were aspects of the information seeking and online searching behaviour. The humanities scholars searched for more named individuals, geographic terms, chronological terms, and discipline terms. This finding has significant implications for developing thesauri and online search aids in the humanities, suggesting that thesauri developed for humanities should incorporate more comprehensive sets of geographic and chronological terms, as well as proper names[25].

Finally, it is important to choose the right data model for publishing your linked data. Terminologies with no structure, such as lists of person names, are not suitable for publishing using the SKOS-scheme. In this case, a model like FOAF-should be envisaged[26].

5.1.2 Is the terminology multilingual?

[22] ANSI/NISO Z39.19-2005, p. 136
[23] Miller U. (2003)
[24] Morville P. et al. (2007)
[25] Shiri A. (2012), p. 75
[26] FOAF (Friend of A Friend) is more convenient for list of names: http://www.foaf-project.org/

As the Athena-survey from 2009 has shown (see supra), thesauri are principally developed in one language. However, multilingual thesauri can play a significant role in facilitating cross-cultural communication in an increasingly globalized information society[27].
Experiments (e.g. with manual mapping of concepts in xTree) have shown that multilingualism is a great asset when mapping several controlled vocabularies. It would be interesting to consider one (international standardized, centrally managed) thesaurus as a "reference"-terminology. Other local terminologies in various languages can then be mapped to this reference terminology.

It is interesting to look at the results of the multilingual terminology mapping experiments at Europeana done within the framework of EuropeanaConnect, as was presented by Ms. Vivien Petras at the *Seminar on Multilingualism and Terminology* in Paris this year[28]. The Europeana online database aggregates data from many European countries. Some data in Europeana are language independent, such as images. But metadata are language-dependant, and by consequence multilingual. This causes retrieval problems, since identical concepts in different languages are not linked.

The Europeana Data Model (EDM), used for aggregating data from content providers, allows alignment of vocabularies via a "semantic data layer", using SKOS or RDF-triples. An experiment was done aligning to pivot vocabularies, e.g. UDC, DDC, VIAF,TGN, Geonames, Wordnets and dbPedia with the property *skos:exactMatch*. The methodology consisted of conversion to SKOS/RDF, different alignment methods (lexical matching, structure-based matching and instance-based matching), disambiguation of matching candidates and combining alignments. The tool used was Amalgam (see supra). More than 500.000 mappings in English, French and Dutch could be realized. Other projects discussed were the European Library and the MACS Initiative, where Library of Congress Subject Headings (LCSH), Rameau and SWD subject headings were mapped. And Europeana 1914-18, a project where keywords were translated in 8 languages and mapped to LCSH.

An experiment with automatic multilingual enrichment in Europeana, mapping several Dublin core elements to the GEMET-thesaurus, DBpedia, Semium Time Ontology and GeoNames saw some unfortunate results due to vocabulary ambiguity: at some point the country "India" was mapped to the concept "poison", because the French "Inde" is the translation or a type of "poison" in another language.

Europeana now wants to evolve to a linked data-based Europeana Data Model, where links to contextual vocabularies from providers are made in the production system (of the provider), as well as enrichment during the ingestion phase. More ideas to solve the language problem in Europeana include improved heuristics of enrichment and stricter normalization, metadata annotation through user input (e.g. social tagging), Geoparser and Gazetteer for creation of geographical data based on place names, open ontology for named periods to use in enrichments and extended enrichments of Agents and Concepts based on DPpedia.

5.1.3 Do concepts have identifiers?

Berners-Lee outlined a set of rules for publishing data on the web in such a way that all published data become part of a single global data space, one of which was the use of *uniform resource identifiers* or URI's.
Berners-Lee speaks of *linked data principles*:

1. Use URIs as names of things
2. Use HTTP URIs so users can look up those names
3. When someone looks up a URI, provide useful information, using the standard (RDF, SPARQL)
4. Include things to other URIs, so that users discover more things

URI's are references to digital objects. These objects can be images, texts, movies, but also metadata-records in a collection management system. There are two types of URIs. A URL (Uniform Resource Locator) is an identifier of the place where something is located and a URN (Uniform Resource Name) give the record a fixed name[29]. The URI's should also be persistent identifiers.

[27] Shiri A. (2012), p. 205
[28] V. Petras (2013)
[29] Glossarium CEST: www.projectcest.be

Concepts that are already published and have URI's mean that they are reusable on the web and can be linked to other concepts. The URI's should be used and persistent when imported in the TMP. When terminologies do not have URI's, they should be given a new URI in the TMP. It is strongly advised that URI's are *recycled*, e.g. the unique numerical codes of concepts in the Art and Architecture Thesaurus can also serve in smaller terminologies, to enhance interoperability.

An important issue to be addressed here is the notion of provenance of URIs. URIs are ideally assigned to concepts by the terminology authority. This ensures the persistency of the URIs when concepts are shared and reused on the web. When the TMP assigns (new) identifiers to concepts, it is important that the TMP can guarantee that these URIs are persistent and that they will remain so when the project has ended.

5.1.4 Is the terminology free of rights?

In a first phase, the terminologies will be imported in the TMP and mapped to concepts in other terminologies. In a second phase, this "meta-thesaurus" can be exported in SKOS/RDF for reuse on the web or as an enriched terminology in the local database.

If we want to reuse the terminology and link it to other sources, it is important that the terminologies are free of rights. As you will see below, a number of terminologies have been published under a Creative Commons Licence, standard arrangements of public copyright license, which allows sharing and re-use of metadata.

If the TMP implements open as well as copyright-licensed terminologies, we should consider an inclusive version, containing all terminologies, as well as a public registry of terminologies available for reuse.

5.2 Survey Questions

The questionnaire is based on some older questionnaires from other projects or cultural organisations, among those are:

- *Identification of existing terminology resources in museums*, ATHENA (see supra)
- *Inventarisatie Terminologiebronnen*, DEN (see supra)

The questionnaire was structured in broad categories, each containing several questions relating to: *Organization*, *Terminology*, *Management*, *Standards* and *Export formats*, *Target Groups*, *Accessibility*, *Rights and Semantic Enrichment*:

Category	Questions	Explication
1. Organization		
	Name of organization	Name of project partner
	Person responsible for the terminology	Name of person managing the terminology
2. Terminology		
	Terminology title	Title of terminology (if available)
	Terminology type	Type of terminology: thesaurus, flat term list, classification, formal ontology
	Provenance of the terminology	Was the terminology bought, developed in-house or adopted from another source?
	Domain	e.g. painting, architecture, literature etc.

	Predominant use of terminology	In house database, online catalogue, scientific document etc.
	Number of terms	How many terms/concepts does the terminology contain ?
	Types of relations	Narrower/Broader, Preferred/Alternatice, Related
	Scope notes	Are the concepts provided with scope notes?
	Multilingualism	Is the terminology multilingual? If so, in how many languages?
3. Management		
	Responsibility	Is there a person responsible for the terminology. If yes, is this a fulltime activity?
	Adding, deleting, changing concepts	Is there a committee of council in your organization where is decided whether concepts will be accepted or not etc.?
	Compulsory information	Are there any compulsory fields to be filled in for each new term?
	Unique identifier	Does each concept have a unique identifier?
	Rules for adding, adjusting or deleting	Are there rules to follow when adding, adjusting or deleting terms
	Number of organization using you terminology	How many organizations use your terminology?
4. Standards and export formats		
	Thesaurus standards	Was the terminology developed by international standards (ISO, ANSI etc.)
	Export format	In which format can the terminology be exported (XML, RDF, CSV etc.)?
5. Target groups		
	Target group	Who uses the terminology (museum staff, public, library etc.)
	FAQ	Is there a manual or FAQ available?
	Suggesting concepts	Is it possible for visitors to suggest new terms?
6. Accessibility		
	Consulting	How can the terminology be consulted (online catalogue, paper, local network etc.)?
	Online	If the terminology can be consulted online: is it consultable in its entirety or only the linked concepts
7. Rights		
	Terminology rights	Does your organization hold the rights to your terminology?
		Is the terminology free of rights to use and reuse?
	Rights arrangement	Who has arranged the rights to your

		terminology?
	(N/A)	My organization has not yet thought about this issue
8. Semantic enrichment		
	Mapping	Is it possible to map your thesaurus via a thesaurus management tool? If yes, with which tool?
		Have you already mapped your vocabulary to another source?
	SKOS	
		Are you familiar with SKOS?
	Linked data	Do you intend to link your terminologies to others in the future? Why/Why not?
	Drawbacks	Do you find there are drawbacks or obstacles that keep you from making public your data (e.g. as LOD? E.g. lack of time/money, loss of control, copyright issues etc.)

5.3 Participating organizations

The questionnaire was sent to all Linked Heritage and AthenaPlus partners. For a list of partners please look at www.linkedheritage.eu and www.athenaplus.eu

Organisations who filled in the questionnaire:

- Koninklijke Musea voor Kunst en Geschiedenis (*Royal Museums of Art and History*), Belgium (KMKG)
- Ministère de la culture et de la communication, France (MCC)
- Stiftung Preussischer Kulturbesitz, Germany (SPK)
- Philipps Universitaet Marburg, Germany (UNIMAR)
- Petőfi Irodalmi Muzeum (*Petőfi Literary Museum*), Hungary (PIM)
- Biblioteca nazionale centrale di Roma, Italy (BNCRM)
- Lietuvos Dailes Muziejus (*Lithuanian Art Museum*), Lithuania (LAM)
- Riksarkivet (*Swedish National Archives*), Sweden (RA)
- Šiauliu Aušros Muziejus (*Šiauliu Aušros Museum*) , Lithuania(SAM)
- Muzej za umjetnost i obrt (*Museum of Arts and Crafts*), Zagreb, Croatia (MUO)
- Ayuntamiento de Girona (*City Council of Girona*), Spain (SGDAP)
- Association Européenne pour la Culture Juive, France (EAJC)[30] :
    - Ben Uri Gallery-London Jewish Museum of Art (EAJC-BU)
    - Hungarian Jewish Archives (EAJC-HJA)
    - Medem Library (EAJC-ML)
    - Rosenthalia Library (EAJC-RL)
    - Steinheim Institute (EAJC-SI)
    - Jewish Museum Prague (EAJC-JMP)
- Biblioteca Academiei Române (*Romanian Academy Library*), Romania (BAR)
- Central Library of the Bulgarian Academy of Sciences (CL-BAS), Bulgaria
- Koninklijk Instituut voor het Kunstpatrimonium (*Royal Institute for Cultural Heritage*), Belgium (KIK-IRPA)
- Museum of Fine Arts – Hungarian National Gallery, Budapest, Hungary (MNG)

[30] The Jewish Museum of Prague and the Hungarian Jewish Archives do not use terminologies. The Rosenthalia Library did not respond to the questionnaire yet.

- Institut Umeni - Divadelni ustav (*Arts and Theatre Institute*), Czech Republic (IU-DU)
- Gottfried Wilhelm Leibniz Universitaet Hannover, Germany (LUH)
- Livrustkammaren och Skoklosters slot med Stiftelsen Hallwylska Museet (*The Royal Armoury, Hallwyl Museum, Skokloster Museum*), Sweden (LSSHM)
- Ministero dei Beni e delle Attività Culturali e del Turismo (MiBAC), Italy
- Nationalmuseum Stockholm, Sweden (NMS)
- Rijksbureau voor Kunsthistorische Documentatie, The Hague, The Netherlands (RKD)
- Estonian Ministry of Culture (EVK)

To this list we add some organisations who will provide their terminology for mapping experiments in the TMP but did not participate in the terminology questionnaire. These organisations are content providers, independent collaborators or projects which signed a memorandum of understanding with Linked Heritage of AthenaPlus. Some of them have already uploaded and mapped in the thesaurus management tool xTree:

- Plantin-Moretus Prentenkabinet, Belgium (PMPK)
- Stedelijke Musea Mechelen, Belgium (SMM)
- Vocabulary of the Israel Museum, Jerusalem (IM)
- PACTOLS-thesaurus, Frantiq-CNRS, Paris, France (FRANT)
- AICIM-thesaurus, Fédération Wallonie-Bruxelles (FWB)
- EU Photography-thesaurus (EUPHOTO)

5.4 Resources : terminologies

The filled-in questionnaires contain information about one or more terminologies. These terminologies can be in-house terminology lists used by just one organisation - or they can be adopted (or bought) from an international standard, e.g. GND from the German National Library, and consequently used by many different organisations. This means that it is possible that terminologies occur more than once in the summary.

Organisation	**Terminology**
SPK	SYSTEMATIK des Museums Europäischer Kulturen (*SYSTEMATIK of the Museum of European Cultures)*
SPK	GND (Integrated Authority File) of the German National Library
LUH	GND (Integrated Authority File) of the German National Library
LUH	STW Thesaurus for Economics
NMS	Geografi-thesaurus
NMS	Kategori-thesaurus
NMS	Material-thesaurus
UNIMAR	Material terms
UNIMAR	Technique terms
UNIMAR	Object type terms
UNIMAR	Classification terms
MCC	Thésaurus de la désignation des œuvres architecturales et des espaces aménagés
MCC	Thésaurus-matières pour l'indexation des archives locales
RKD	Getty AAT (*Dutch version*) – Getty Research Institute
RA	Thesaurus on cartography and historical and archival research

PIM	Authority file: Hungarian Biographical Index
PIM	Authority file: Hungarian Emigrant Writers
PIM	Authority file: Bibliography of Contemporary Authors
PIM	Authority file: Awards and Awardees
PIM	Authority file: Hungarian Recipients of the Legion of Honours
PIM	Authority file: Graduate database
PIM	Authority file: The Hungarian Nobility Genealogy
PIM	Authority file: Hungarian Family History Repository
PIM	Authority file: Budapest Topography
PIM	Authority file: Graves of Hungarian Writers
PIM	Subject Headings (current) – Köztaurusz (in near future) (only the last one is considered in this survey)
LSSHM	Classification-thesaurus of the Royal Armoury, Hallwyl and Skokloster museums
BAR	Numismatic and sphragistic terminology
SAM	Classification systems for museums and museum objects (LIMIS-thesaurus)
LAM	Classification systems for museums and museum objects (LIMIS-thesaurus)
KIK-IRPA	KIK-IRPA thesaurus (multiple domains)
KIK-IRPA	KIK-IRPA persons and institutions
KIK-IRPA	KIK-IRPA Christian faith thesaurus
MUO	Thesaurus of the Museum of Arts and Crafts, Zagreb
BNCRM	Nuovo Soggettario
BNCRM	Dewey Decimal Classification (DDC)
MNG	Keywords-thesaurus of the Museum of Fine Arts-National Gallery, Budapest
CL-BAS	Universal Decimal Classification (UDC)
EUPHOTO	Europeana Photography Vocabulary
IU-DU	Universal Decimal Classification (UDC)
IU-DU	Theatre Institute Library Classification of Specialized Literature (TILC)
KMKG	KMKG/MRAH-Object name thesaurus
KMKG	KMKG/MRAH-Geography thesaurus
KMKG	KMKG/MRAH-Techniques and Materials
SGDAP	Descriptors-LEMAC (thesaurus of the City Council of Girona) – Catalan adaptation of LCSH
SGDAP	Descriptors- names (people, places, organisations) (thesaurus of the City Council of Girona)
MiBAC	Thesaurus PICO 4.3
EAJC-BU	Getty Art and Architecture Thesaurus – Getty Research Institute
EAJC-ML	Subject headings
EAJC-SI	Getty Art and Architecture Thesaurus – Getty Research Institute
PMPK	Vocabulary *Plantin-Moretus prentenkabinet*

SMM	Vocabulary *Stedelijke Musea Mechelen*
BT	British Museum Thesaurus
IM	Vocabulary Israel Museum
FWB	AICIM-thesaurus
FRANT	PACTOLS-thesaurus
EVK	MulS sõnastikud

A total of 57 terminologies could be identified. When we leave out the terminologies that are recorded twice we can distinguish 52 individual terminologies:

- The UDC-classification is used by two organisations
- The Art and Architecture Thesaurus is used by 3 organisations (when we consider the Dutch version of the RKD as identical as the American version)
- The GDN of the German National Library is used by two organisations
- The thesaurus created for LIMIS is used by two organisations

The results of the terminology questionnaire are discussed below. Important remark: due to the variety in answers in the questionnaire, it is not possible to deduce exact percentages. The results are presented as general outlines.

5.5 Survey results

5.5.1 Type of terminology

Type of terminology

In this section we look at what kind of terminology is used in the institution. The terminology can be a thesaurus, a flat term list, classification or a combination of types.

- 40 of the terminologies could be identified as thesauri, i.e. a structured vocabularies with at least a hierarchical relation between terms (broader/narrower relation). Most of the terminologies, such as the KMKG-thesauri and the LIMIS-thesauri contain only hierarchical relations. Other thesauri, such as the AAT and the Nuovo Soggettario, also contain equivalence and associated relations.
- The SYSTEMATIK-classification used by SPK, and the Universal Decimal Classification used by The Arts and Theatre Institute in Prague and the Central Library of the Bulgarian Academy of Sciences, are classification systems.
- The authority files of the Petőfi Literary Museum and the names-list of the City Council of Girona is a names-list. The vocabulary used by the Museum of Arts and Crafts in Zagreb is also a flat term-list.
- A number of terminologies use a combination of flat lists and thesauri or classification systems. This is the case for e.g. the terminologies of UNIMAR, where the relations between terms are not consistently applied. The MulS-vocabulary of the Estonian Ministry of Culture combines several different types of terminologies, including flat term lists and thesauri.

Scope notes

A scope note is a short description following a term explaining its coverage, specialized usage, or rules for assigning it.

The terminologies do not often provide terms with scope notes. 8 thesauri (Art and Architecture Thesaurus, the Universal Decimal Classification, PICO 4.3, Nuovo Soggettario, MCC-thesauri, one authority list of the Petőfi Literary Museum and the national Hungarian standard Közstaurusz) contain scope notes.

Other terminologies, the majority, have no scope notes at all or have some, such as in the thesauri of KIK-IRPA, PACTOLS-thesaurus and the thesauri of the Swedish national archives.

Thesaurus development

This question assesses the origin of the terminology. There are two possibilities:

- ***In-house development***: this means that the terminology was created by staff members of an organisation specifically designed for the collection. Sources of the terminologies can be national or international standards.
- ***Adopted / bought***: this means that the terminology was taken over in its entirety from another (specialized) organization. It is possible that the terminology is freely available, e.g. downloadable from a web service; or was purchased from a (specialised) organisation.

Only 6 terminologies were adopted or bought, more precisely the GND of the German National Library, the AAT, the Nuovo Soggettario, the Kösztaurusz, the UDC and DDC. The UDC is a classification system in 51 languages and a global, widely spread classification systems, mainly used for classification in libraries and archives (it is not a multilingual thesaurus). The DDC is of the same type and is used in 200,000 libraries in at least 135 countries.

The AAT is probably the most renowned international thesaurus for concepts in domains of art and architecture. It can be implemented in e.g. an Adlib-database, but it is currently not free of rights (only single concepts, not complete structures can be reused) or exportable from a web service, so we can assume that the thesauri were bought. National standards, such as the GND for indexing in the library sector or the recent Hungarian Köstaurusz are national standards which are (or are to be) used by many institutions in a certain country.

However, the majority of partners developed their terminology in-house, with specific metadata schemes and thesauri “tailor-made” for their own collections. This doesn’t mean that they are not compliant with international standards: e.g. several partners noted that for geographic names references are made to the Getty Thesaurus Geographic Names (TGN).

Domain

The organizations who filled in the questionnaire can all be situated in the GLAM-sector (galleries, libraries, archives and museums). A division between specific domains (e.g. arts and crafts, painting, literature) could not be done because the metadata often overlap.

The thesauri contain terms or concepts correlating with following metadata:

- Object names / keywords
- Person names
- Geographical locations
- Organisation names
- Materials
- Techniques
- Iconography
- Style/period
- Collection type

Predominant use

This question investigated the main purpose or goal of the terminologies. We suggested three main uses:

- As an indexing tool for an in-house database
- As an search query feature for an online catalogue
- As a scientific document

Almost all recipients use the terminology both as an indexing tool and a search query tool for an online catalogue or collection. The SYSTEMATIK-classification and GND used in SPK, the flat term list used in the Arts and Crafts Museum in Zagreb and the thesaurus of the Museum of Fine Arts – National Gallery in Budapest are only used in a local database or management system. The numismatic and sphragistic terminology of the Romanian Academy Library is considered primary a scientific document.

The Europeana Photography Vocabulary is used in the MINT-tool where it is mapped to EDM.

Number of concepts

The terminologies contain between 70 (Museum of Fine Arts, Budapest) and more than one million concepts (GND).

Multilingualism

With this question we wanted to learn more about the language of the terminologies. Mapping experiments have shown that it is a great advantage when concepts are multilingual.
Even though some of the international standards have stimulated multilingualism, such as the AAT and DCC, terminologies constructed in-house are mainly developed in the language of the creator. In this survey only the terminologies listed below are multilingual, even though not always completely:

- TILC-thesaurus in Czech, English, German
- LIMIS-thesaurus in Lithuanian, English, Russian, French, Latin etc. (not consistent)
- KMKG-thesauri in French, Dutch, English
- Thesauri National Museum in Stockholm, Swedish, English (not consistent)
- KIK-IRPA thesauri in Dutch, French (not consistent)
- AAT in Dutch, English, Spanish and partly in French, Portuguese, Chinese and German
- PICO 4.3 in Italian and English
- EU Photography-thesaurus in English, Italian, German, French, Dutch, Danish, Catalan, Spanish, Bulgarian, Polish, Slovene, Lithuanian (Mandarin and Chinese in progress)
- PACTOLS-thesaurus in French, Spanish, Dutch, Italian and German
- Vocabulary of the Israel Museum in English and Hebrew

5.5.2 Management

Thesaurus management

The questions listed in this part of the survey investigate the management of the terminology. Is there someone responsible for the terminology and if so, how much time does this person spend on the management of the thesaurus on a weekly basis? Management includes: updating, adding, replacing and deleting terms, writing scope notes, translating, mapping etc. We were also interested in creating a list of names of thesaurus managers, in order to allow good communication when starting to work in the TMP.

International or national standards such as AAT, UDC, GND etc. have full-time or part-time thesaurus managers. For the Dutch AAT, managed by the RKD in The Hague, one person spends ca. 32 hours a week managing the thesaurus. The Nuovo Soggettario is managed by a central service located in the

National Central Library of Florence. The Estonian Ministry of Culture engaged one person half-time for thesaurus management.

The Romanian Academy Library, KIK-IRPA, the Museum of Arts and Crafts in Zagreb and the Arts and Theatre Institute in Prague have no specific thesaurus manager in their service working on their in-house vocabularies. KIK-IRPA says that the thesaurus is a shared responsibility of multiple people working with the collection management system. The Museum of Fine Arts – Hungarian National Gallery in Budapest will engage a full-time thesaurus manager in the future, but now a museum professional is dealing with in-house projects such as the thesaurus.

The UNIMAR-vocabularies are managed by 5 staff members: adding new terms, aligning requirements of various cataloguing projects, extending vocabulary files and correcting cataloguing data. The City Council of Girona has three technicians working on the vocabularies. The KMKG-thesauri are centralised and managed by one staff member.

The authority files used by the literary museum in Budapest are managed by two people. Sometimes more than one organization manage the thesauri: the LIMIS-thesauri are edited by the Lithuanian Art Museum, with contributions from the Siauliai Ausros Museum.

Even though most organizations have thesaurus managers, the time spend on the terminologies (when created and managed in-house) is not regarded as a full-time activity.

UNIMAR, has two staff members working 50% exclusively for vocabulary control (including authority data control) and three persons in varying shares, dependent on the projects and tasks.

We also informed about the existence of a committee or “thesaurus working group” in the organization, who meet on a regular basis for discussing thesaurus topics. 31 recipients who answered this question, said that they don’t have such committees and 12 said they did. Some organizations do not have a regular thesaurus committee, but work with ad-hoc working groups, experts or staff members. The City Council of Girona explains that anyone who is documenting can add new terms. These terms are revised and accepted or deleted afterwards, when the work is completed.

The Museum of Fine Arts in Budapest installed a professional working group consisting of museum professionals, collection managers and one professional coordinator. For the Nuovo Soggettario, a national committee is responsible for decisions encompassing thesaurus work. Finally, the Estonian Ministry of Culture installed a Terminology working group.

Thesaurus Rules and Procedures

27 organisations claim they have specific procedures or rules for adding, changing or deleting thesaurus terms. 14 answered they do not have such rules. The PICO 4.3 thesaurus for example, implements following procedure: “*any modification of the thesaurus must be approved by the maintenance team. Moreover, specific rules must be followed for adding or modifying terms, for moving terms within the hierarchy, and to assign URIs to terms in the different versions of the thesaurus*”.

Vocabularies are also controlled by limiting the persons adding, adjusting or deleting concepts. At the Estonian Ministry of Culture and KMKG, for example, only one person has access to the thesaurus module of the collection management system.

Uniform Resource Identifiers (URIs)

All organisations, except for one, work with a collection management system. Only the Romanian Academy Library does not yet have a collection management software, but is planning on purchasing one in the future.

39 terminologies have concepts with URI’s, 11 terminologies do not have URI’s. Other terminologies have partial or no URI’s. The reason for this is that some software programmes do not provide easy-to-access identifiers for concepts, such as FilemakerPro or MuseumPlus. MuseumPlus has a unique numerical code for each concept, but these codes are only stored “in the back-end”. They can be retrieved when exporting in XML, but this feature is not always available in MuseumPlus. For example, only metadata-records of the KMKG have persistent URI’s.

Some organisations have concept with new URI's which were created when they were imported in de xTree-tool. An example is the vocabulary of the Plantin-Moretus Prentenkabinet.

Terminology use

We asked how many use the terminology which is used by the organizations. In 13 cases, in-house terminologies are only used by the organization itself. The UNIMAR-terminologies are used by 12 affiliated institutions, the LIMIS-thesauri by ca. 60 organizations in Lithuania. Also the Muls-thesaurus of EVK is used by ca. 60 organisation in Estonia.

The thesauri of the Council of Girona is used by the council itself, but it is not certain other organization used it also since the terminologies are published.

The Dutch AAT is used by over a hundred institutions in The Netherlands and Belgium. The Bulgarian UDC is certainly used by the 49 libraries affiliated to the Central Library of the Bulgarian Academy of Sciences. All public libraries that want to contribute to the Union Catalogue of Czech Republic are required to use the Czech UDC.

The PICO 4.3 thesaurus is used by all central and local Italian institutes depending from MiBAC and the Nuovo Soggettario is used by all organizations to the National Library Service (BNCRM). The GND is used by "some hundred" libraries and archives in Germany. The PACTOLS-thesauri and the AICIM-thesauri are used by several cultural organizations they represent.

Standards and Export Formats

In order to have an overview of the different formats that we will work with for importation in the TMP, we asked the organizations to sum up the formats in which the terminologies can be exported from their management systems.

43 terminologies can be exported in XML, e.g. the authority files of the Literary Museum in Budapest can be exported using MARCXML, the UNIMAR-thesauri using HiDA4 Bildarchiv and the thesauri of the National Museum of Sweden can export an XML format from MuseumPlus.

The Dutch AAT is available in XML via the Adlib-software and in SKOS via a webservice.

Large thesauri with centralized management are often already available in SKOS/RDF: UDC, Nuovo Soggettario, the vocabulary of the EU Photography-project, the thesauri of the MCC (using GINCO-software), the Hungarian köztaurusz, PICO 4.3, the GND of the German National Library and the PACTOLS-thesaurus of the CNRS (OpenThésau software). The Getty AAT (American version) will be published as linked open data soon.

Some thesauri, such as the Israel Museum-thesaurus, the thesaurus of the Stedelijke Musea Mechelen and Prentenkabinet Plantin-Moretus can be exported in SKOS/RDF from the xTree-tool where they were uploaded and mapped.

The thesauri developed for the in-house LIMIS-software (Lithanian Museum Integral Information System), the thesauri of KMKG and Museum of Fine Arts in Budapest, the thesauri of the City Council of Girona (Fotostation), the AICIM-thesaurus of the Fédération Wallonie-Bruxelles (FilemakerPro) and the terminology of the Estonian Ministry of Culture are currently exportable in CSV-format.

The in-house database ProMus of the Arts and Crafts Thesaurus of Zagreb and the vocabulary of the Romanian National Academy are not (yet) exportable. The thesauri of the Swedish National Archives use in-house CMSs or text editors, but have no specific export functions.

Target Groups

25 terminologies aim at staff members and visitors of library and archives. 36 terminologies are directed to staff members and visitors of museums.

5.5.3 Accessibility

This information investigates how the vocabularies can be consulted: online, in a local database or on paper?

The UNIMAR-thesauri, the SYSTEMATIK-classification used by SPK, the thesaurus of the Hungarian Museum of Fine Arts-National Gallery, the thesaurus of the Museum of Arts and Crafts of Zagreb, the terminology of the Swedish national archives and the terminology of the Estonian Ministry of Culture are only consultable in a local database and/or on paper.

Terminologies listed in the table below can be consulted online, even though not always entirely (e.g. often only linked terms):

Terminology	**URL**
Thésaurus de la désignation des oeuvres architecturales et des espaces aménagés (MCC)	http://data.culture.fr/thesaurus/page/ark:/67717/T96
Thésaurus-matières pour l'indexation des archives locales (MCC)	http://data.culture.fr/thesaurus/page/ark:/67717/Matiere
Category, geographic origin, material (NMS)	http://www.nationalmuseum.se/samlingarnaonline
Classification systems for museums and museum objects (LIMIS-thesaurus) (SAM)	www.limis.lt
Classification systems for museums and museum objects (LIMIS-thesaurus) (LAM)	www.limis.lt
Royal Armoury, Hallwyl museum, Skokloster museum (Sweden)	emuseumplus.lhs.se
Descriptors-LEMAC + descriptors (thesaurus of the City Council of Girona)	
GND (SPK and LUH)	http://d-nb.info
Dutch AAT	http://browser.aat-ned/nl
Authority lists (PIM)	http://opac.pim.hu/index.jsp?page=search&group=1
Numismatic and sphragistic terminology	http://www.biblacad.ro/ratonline.html
KIK-IRPA thesauri	http://www.kikirpa.be
Nuovo Soggettario (BNCRM)	http://www.sbn.it/opacsbn/opac/iccu/base.jsp

UDC (CL-BAS)	http://aleph.cl.bas.bg/F/9MJP1EE6C6PINUJVFR6MQBFJJVI1X8Y3T1Q7NV9RRLBLEY735P-00247?func=file&file_name=base-list
Europeana Photography	http://bib.arts.kuleuven.be/photoVocabulary/30100
UDC (IU-DU)	http://aip.nkp.cz/mdt
Theatre Institute Library Classification of Specialized Literature (IU-DU)	http://www.idu.cz/cs/systematicke-trideni-odborne-literatury-v-knihovne
KMKG-thesauri	www.carmentis.be
PICO 4.3 (MiBAC)	http://purl.org/pico/thesaurus_4.3.0.skos.xml
AICIM-thesaurus (FWB)	www.aicim.be
PACTOLS (FRANT)	http://pactols.frantiq.fr/opentheso/
Gett AAT (EAJC)	http://www.getty.edu/research/tools/vocabularies/aat/

5.5.4 Rights

In this section we investigate the rights that are imposed on the terminologies.

SYSTEMATIK-classification (SPK)	Not free of rights (arranged internal)
Dutch AAT (RKD)	Not free of rights (arranged by Getty)
Authority files (PIM)	Not free of rights
Köztaurusz (PIM)	Not free of rights
KIK-IRPA thesauri (KIK)	Not free of rights
DDC (LUH)	Not free of rights
Getty AAT (EAJC)	Not free of rights
PACTOLS-thesaurus	http://creativecommons.org/licenses/by-nc-nd/2.0/fr/
GND of the Deutsche Nationalbibliothek (German National Library) (SPK, LUH)	http://creativecommons.org/publicdomain/zero/1.0/
Thésaurus de la désignation des oeuvres architecturales et des espaces aménagés (MCC)	http://creativecommons.org/licenses/by-sa/2.0/
Thésaurus-matières pour l'indexation des archives locales (MCC)	http://creativecommons.org/licenses/by-sa/2.0/
Thesaurus on cartography and historical and archival research (RA)	Free of rights (arranged by Swedish National Law)
Numismatic and sphragistic terminology	Free of rights

(BAR)	
Classification systems for museums and museum objects (LAM, SAM)	Free of rights (arranged by Lithuanian National Law)
Nuovo Soggettario (BNCRM)	http://creativecommons.org/licenses/by/2.5/
Europeana Photography Vocabulary	Free of rights
Universal Decimal Classification (UDC)	http://creativecommons.org/licenses/by-sa/3.0/
KMKG-thesauri (KMKG)	Free of rights
City Council of Girona-thesauri	Free of rights
Estonian Ministry of Culture (EVK)	Free of rights
Thesauri Nationalmuseum, Sweden	Free of rights
Thesaurus PICO 4.3 (MiBAC)	http://creativecommons.org/licenses/by/2.5/it/legalcode
STW-thesaurus for economics (ZWB)	http://creativecommons.org/licenses/by-nc-sa/3.0/de/
UNIMAR-thesauri	*Organisation hasn't arranged rights yet*
Museum of Arts and Crafts Zagreb-thesaurus	*Organisation hasn't arranged rights yet*
Museum of Fine Arts, Budapest, keywords thesaurus	*Organisation hasn't arranged rights yet*
TILC-thesaurus (IU-DU)	*Organisation hasn't arranged rights yet*

5.5.5 Semantic Enrichment

We asked the respondents these questions:

- Is the terminology used by the organization already mapped to another terminology?
- Are you familiar with SKOS?
- Has the terminology already been published as LOD?

Some terminologies have been mapped to other sources or they have experimented with mapping:

- The GND of the German National Library is partly mapped to VIAF (Virtual International Authority File)
- EU Photography thesaurus is mapped to the PartagePlus-thesaurus
- The *Thésaurus de la désignation des oeuvres architecturales et des espaces aménagés* (MCC) is mapped to other in-house source with the GINCO-tool
- *Thésaurus-matières pour l'indexation des archives locales* (MCC) is mapped to RAMEAU and Dbpedia
- The German *Thesaurus for Economics* is mapped to the Eurovoc-thesaurus
- The Israel Museum thesaurus is mapped to the British museum thesaurus in xTree
- The KMKG-thesauri are mapped to British Museum thesaurus in xTree
- The National Archives of Sweden have made mapping experiments with in-house terminologies.

Most organizations have notion of SKOS, even though not on an experienced level.

The PICO-thesaurus is available in LOD and exposes the thesaurus through a SPARQL endpoint within dati.culturaitalia.it, at:
http://dati.culturaitalia.it/sparql/browse?describe=%3Chttp%3A%2F%2Fculturaitalia.it%2Fpico%2Fthesaurus%2F4.3%3E

PSK is not planning on linking the SYSTEMATIK-classification to other sources, because they consider it an in-house instrument. Other organisations are not willing to link because they do not own the rights (e.g. adopted terminologies) or the terminology is not considered free of rights. Others want to review their thesauri before linking them (National Museum Sweden).

However, most organisations are planning on linking to other terminologies. Some interesting statements and drawbacks are listed below:

- *Lack of time and resources. We would like to review our terminologies for our own sake ("Kategori"), but especially before making it public which seems an impossible project with the few resources we have in the field of Digital Documentation. However, the "Kategori" terminology was developed alongside collection digital management and it would be very useful for us to see our terms in a broader context. The geographical references provide an interesting possibility of linking objects and information* (Nationalmuseum, Sweden)
- *Lack of time and money* (Petőfi Literary Museum, Budapest)
- *Lack of time* (Estonian Ministry of Culture)
- *Lack of money, time, personnel, copyright issues* (Šiauliu Aušros Muziejus)
- *First we need an advanced terminology (thesaurus, developed by international standards); we want to make our objects part of the European digitized heritage and ease retrieval on the international level. Lack of money, lack of personnel is a drawback. Participation in EU-projects can help to learn* (Museum of Arts and Crafts Zagreb)
- *This is the key to be more accessible and improve interoperability. We will map our vocabulary to a SKOS vocabulary for the Europeana Photography project. We are also implementing new software to manage digital objects, add metadata and enable a new web service for consulting digital objects* (City Council of Girona)
- *Lack of time can be a drawback, also control issues need to be clarified. Mapping is extremely time intensive* (KIK-IRPA)
- *We are developing a new website, so we need some time and money to implement it* (Museum of Fine Arts, Budapest)
- *There are no drawbacks, Linked Data and data enrichment are a strategic objective for MiBAC-ICCU* (MiBAC)
- *Lack of online resources* (National Archives of Sweden)
- *Lack of resources to do it properly* (UNIMAR)
- *The Strategy of Digitization 2013-2020 [in Croatia] is presently finalized and it includes creation of a common thesaurus for Croatian museums according to international standards, common repository and aggregator* (Museum of Arts and Crafts, Zagreb)

6 Suggested strategy for import

The information collected from the questionnaires can serve as a basis for tracing the outlines of a registry planning. In November, a production version of the TMP will be available. By then, the registry of terminologies must be stored in the TMP and ready for mapping.

It would be advisable that some international standards (preferably in English) are imported in the TMP, containing many (multilingual) concepts and covering the broader cultural sector, such as the AAT and VIAF or UDC. These thesauri can serve as “reference”-thesauri, where smaller, monolingual can be mapped to.

In the summary below I have made a distinction between terminologies that are ready for import in the TMP, terminologies that are not yet ready for import in the TMP and terminologies that are not suitable for import in the TMP.

In WP4 we will encourage as many organizations as possible to import their terminology in the TMP. This includes terminologies from Linked Heritage or other projects. We will contact these organizations and look at the possibilities of an agreement, which would entail the use and reuse of their terminologies in the TMP.

Ready for import

UDC (IU-DU, CL-BAS), in Czech and Bulgarian → subject headings (library domain), CC BY-SA 3.0 (free to share, remix, commercial use, with restrictions)

KMKG for Europeana Photography → object names, materials, techniques relating to photography (already imported in TMP), not free of rights (?), multilingual

Materials, Classification, Techniques, Vocabulary, UNIMAR → art history, archaeology: architecture, painting, sculpture, applied arts, graphic art (rights issue needs to be arranged by institution)

City Council of Girona: LEMAC-thesaurus → subject headings related to Girona (topics, places, places of Girona), free of rights

GND of the Deutsche Nationalbibliothek (SPK, LUH) → geographical information, keywords (certain categories), CCO public domain

PICO 4.3 (MiBAC) → cultural heritage, archaeology, architecture, audio-visual, contemporary art, decorative art, ethnology, fine arts, furniture, geography etc., Italian CC licence (to be translated), SKOS

Authority file: Budapest topography (thesaurus category) (Petőfi Irodalmi Muzeum) → geographical references; export in MARCXML, free of rights

Köztaurusz (Petőfi Irodalmi Muzeum) → (general Hungarian culture), SKOS, already imported in TMP, not free of rights

Nuovo Soggettario (BNCRM) → subject headings (library domain), CC BY-SA 3.0 (free to share, remix, commercial use, with restrictions), multilingual

Thésaurus de la désignation des oeuvres architecturales et des espaces aménagés and Thésaurus-matières pour l'indexation des archives locales (MCC) → architecture ; archives ; CC-licence, SKOS

Classification systems for museums and museum objects (LIMIS-thesaurus) (SAM, SAM) → art, art history archaeology; free of rights, multilingual

Vocabulary Plantin-Moretus prentenkabinet → art history (archives); free of rights, SKOS

Vocabulary Stedelijke Musea Mechelen → art, art history, archaeology; free of rights, SKOS

PACTOLS (Frantiq, CNRS) → archaeology; CC-licence, SKOS, multilingual

Vocabulary Israel Museum (IM) → history, art history, archaeology; SKOS, multilingual

Currently not ready for import

MulS sõnastikud, Estonian Ministry of Culture→ (general Estonian culture), free of rights, no unique identifier

Thesaurus of the Museum of Arts and Crafts, Zagreb → painting, sculpture, architecture, design, applied arts, photography; flat terms list, export possibilities from in-house CMS not specified in questionnaire

Numismatic and sfragistic terminology, Romanian Academy Library, Bukarest → numismatics, sphragistics, heraldry; flat term list, no CMS used

SYSTEMATIK des Museums Europäischer Kulturen (Museum of European Cultures), SPK → cultural history; no intention to export or link to other vocabularies

Art and Architecture Thesaurus and *Dutch Art and Architecture Thesaurus* (EAJC, RKD) → Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, Objects, Brand Names in domains of art and architecture; rights need to be cleared (wait for LOD-version), SKOS, multilingual

Theatre Institute Library Classification of Specialized Literature (IU-DU) → performing arts; no export possible, rights need to be cleared

Thesauri of Swedish National Archives (RA) →cartography, historical and archival research; no specific export functions, no identifiers

Dewey Decimal Classification (BNCRM) → subject headings (library), rights need to be cleared, multilingual, SKOS

Classification (Royal Armoury, Hallwyl Museum, Skokloster Museum) → painting, weapons, costume etc., free of rights, thesauri need to be standardized

Keywords (Museum of Fine Arts – Hungarian National Gallery) → fine arts; no identifiers, rights need to be cleared

Thesaurus object names, Christian faith thesaurus (KIK-IRPA) → cultural heritage; rights need to be cleared; identifiers in progress (Adlib), multilingual

Thesaurus object names, materials, techniques, geographic locations (KMKG-MRAH) → art, archaeology, applied arts and musical instruments from prehistory until 20th century; free of rights; no identifiers for concepts (MuseumPlus)

AICIM-thesaurus (FWB) → art history, archaeology, applied arts; rights need to be cleared, no identifiers

British Museum-thesaurus (BMT) → art history, archaeology, applied arts; rights need to be cleared, SKOS

Not apt for SKOS (use other model)

City Council of Girona – descriptors containing single names of persons and organizations

GND of the Deutsche Nationalbibliothek (SPK, LUH) → persons, institutions, congresses, titles (subthesaurus)

Authority files (Petőfi Irodalmi Muzeum) → Hungarian bibliographic index, emigrant writers, contemporary authors, awards and awardees, graduate database, Hungarian nobility geneaology, family history, graves of Hungarian writers

Thesaurus person names and institutions, KIK-IRPA → broad cultural heritage

7 Possible future developments

WP4 we can also envisage following activities:

- Encourage organizations that are not partner in AthenaPlus to deliver content to the TMP: e.g. thesauri and vocabularies published as LOD in the SENESCHAL-project (thesauri of English Heritage, Royal Commission on the Ancient and Historical Monuments of Scotland and Wales, University of South Wales), VIAF etc.
- LIDO-actor roles and event types were translated in all partner languages in the Linked Heritage-project. They will be implemented in the MINT-tool in the AthenaPlus-project (it was not possible to do it in the Linked Heritage mapping version). We could continue working on enriching LIDO-types and elements.
- We could consider alignment of the enriched vocabularies in SKOS in the EDM schema, developed by Europeana. This would enhance multilingualism and interoperability in the Europeana online catalogue.

8 Conclusion

This deliverable describes the results of an analyses of GLAM-sector terminologies, the selection criteria used for the collection of suitable reference terminologies and a detailed overview of the collected terminologies.

Some key principles, such as terminology practice and SKOS/RDF and TMP are shortly explained. A state of the art describes previous terminology surveys, terminologies published in SKOS and meta-thesauri. This revealed that some interesting mapping experiments have been conducted, but that mapping of a large number of thesauri (manual or automatic) is still largely unexplored. However, in recent years many vocabularies have been published and made available in SKOS.

A few mapping experiments focus on vocabulary enrichment using semi-automatic mapping tools. Because the results of these experiments - a combination of automatic mapping and manual expert assessment - was mainly positive, WP4 should investigate the possibilities of integration of such (open-source) tools or functions in the TMP. This would save a lot of time and effort when mapping thousands of concepts in the TMP, where this activity is currently done manually.

For the selection of terminologies that will be imported in the TMP, it is important that the type of vocabulary is controlled, that the terminology is reusable (free of rights) and that multilingualism and unique identifiers for each concept in the vocabulary are available. If the TMP assigns a URI to concepts of an imported terminology or a terminology created from scratch in the TMP, the persistency of these URIs must be ensured, e.g. in an agreement or procedure which describes how the TMP will assign and manage PIDs.

The questionnaire collected information about 33 organizations and 57 terminologies. It contains questions regarding *organization*, *terminology*, *management*, *standards* and *export formats*, *target groups*, *accessibility*, *rights and semantic enrichment*.

Most of the vocabularies are of the type thesaurus or classification with at least a hierarchical structure. Some terminologies are flat lists or a combination of several types of terminologies. Scope notes are only available in larger (international) terminologies and only partially or inexistent in most terminologies.

Most of the organizations use their own in-house terminology with no references to standards, but the scope of their use can be quite large, especially when they are developed by centralized organizations on a national level.

Multilingualism is not widespread, only a few terminologies have concepts in more than one language.

Almost all terminologies are well managed, applying specific thesaurus management rules, even though this is not regarded as a full-time activity.

The terminologies are mainly used as an indexing tool and as a query feature in an online database. Most concepts have a unique identifier, even though this largely depends on the CMS which is used. Most terminologies are exportable in a CSV-file from a local database, such as word and excel, others are already available in SKOS/RDF. Only a few terminologies are not exportable. 22 terminologies can be consulted online, but not always completely (often only linked terms).

Concerning the rights imposed on the terminologies, we notice that larger, international or national terminologies apply Creative Commons (CC) licences. Others claim no specific rights and are free to use. However we notice terminologies that are not free of rights and organizations that haven't thought about this issue yet. It would be preferable if terminologies imported in the TMP are licensed under a CC-copyright. We would like to see the topic of rights addressed in AthenaPlus: what kind of licenses are available, what are the possibilities and the consequences for organizations publishing their data under a CC-license, what are the differences between licensing metadata/data and images and scientific notes in metadata etc.

The section concerning rights is also important for the TMP and the possibilities of reusing the enriched data. If we are importing terminologies that are free of rights as well as terminologies which are not, we should think about registration of two mappings: an inclusive master file, encompassing all enriched

terminologies, where the use is restricted to the TMP; and another with those terminologies published under a CC-license. The latter can be used and shared, e.g. in LOD-experiments. We could lay this down in *data use agreements* with providers of terminologies which define the status of their terminology rights: open, under copyright or with an unclear status.

Most organizations are planning to link their terminology to others in the future, but some drawbacks can be a lack of time, personnel, expertise or a lack of qualitative resources. Others wish to revise their terminologies before importing them in the TMP.

Based on the answers in the questionnaire we could trace out a strategy for terminology implementation in the TMP and look at certain gaps and opportunities.

WP4 should try to tackle these drawbacks with more information on rights issues and SKOS. We should also envisage a strong personal approach and guidance when organizations will use the TMP. The mapping procedure is manual, so we could think of ways to set up a procedure which will allow users to map concepts easier. We could also inform partners about open source tools which can be used to automatically "clean up" or process the terminologies before importing them in the TMP, such as *(Google) Open Refine*[31].

It is also very important that organizations that do the effort of importing and mapping their in-house terminology in the TMP will get a tangible return, e.g. an enriched terminology for integration in their local CMS. It is therefore very important that organizations can at least export their own (enriched) terminology in an exchange format such as SKOS/RDF for further reuse by the authority managing the terminology.

Finally there is the provenance issue that should be thought of. When terminologies are imported in the TMP, they should be managed in the TMP, e.g. if a concept is added or changed in the local database, this should be done likewise in the TMP. It would be an idea to set up a procedure or model which provides guidelines to overcome such provenance problems.

[31] http://code.google.com/p/google-refine/

9 APPENDIX 1: REFERENCES

Books and articles:

HODGE 2000
Gail Hodge, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*, Washington 2000.

CAGNOT et al. 2011
Stéphane Cagnot, Johann Holland, Marie-Véronique Leroi, Michael Culture Aisbl, *Your Terminology as Part of the Semantic Web. Recommendations and Guidelines for Design and Management*, Linked Heritage 2011.

DE BOER et al. s.d.
Victor de Boer, Jan Wielemaker, Judith van Gent, Marijke Oosterbroeck, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, Guus Schreiber, *Amsterdam Museum Linked Open Data*, EuropeanaConnect research project, s.d.

HUNTER 2001
Jane Hunter, *MetaNet: A metadata term thesaurus ton enable semantic interoperability between metadata domains*, Journal of Digital Information, 1(8), 2001

MAY et al. 2011
Keith May, Ceri Binding, Doug Tudhope, Stuart Jeffrey, *Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources, United Kingdom 2011*

MILLER 2003
Miller U., *Thesaurus and new information environment*. In: M. Drake and M. N. Maack (Eds.); Encyclopaedia of library and information science, 2nd ed. Boca Raon: Taylor & Francis Group.

MORSHED et al. 2011
Ahsan Morshed, Caterina Caracciolo, Gudrun Johanssen, Johannes Keizer (2011), *Thesaurus Alignment for Linked Data Publishing,* Proc. International Conference on Dublin Core and Metadata Applications 2011

MORVILLE et al. 2007
Peter Morville, Louis Rosenfeld, *Information architecture for the World Wide Web: Designing Large-Scale Web Sites*, 3rd ed. Sebastopol, CA: O'Reilly, 2007

SHIRI 2012
Ali Shiri, *Powering Search. The Role of Thesauri in New Information Environments*, New Jersey 2012

SOERGEL 2003
Dagobert Soergel, *Functions of a Thesaurus/Classification/Onological Knowledge Base*, University of Maryland 2003

Standards, deliverables and presentations:

ANSI/NISO Z39.19-2005, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, 2005

Marie-Véronique Leroi, Johann Holland, **D4.1 : *Identification of existing terminology resources in museums***, Athena (ECP-2007-DILI-517005), July 31 2009.

Vivien Petras, ***Multilingual Terminology Mapping at Europeana***, Berlin School of Library and Information Science, presentation at: *Linked Heritage Seminar on Multilingualism and Terminology*, Paris April 18 2013.

The Vocabulary Mapping Framework (VMF): an introduction, v.1.0, December 12, 2009, developed with funding from the Joint Information Services Committee (JISC), currently hosted and administered by the International DOI Foundation (IDF) under the guidance of an independent multi-stakeholder Advisory Board: http://www.doi.org/VMF/

Websites:

Amalgame (AMsterdam ALignment GenerAtion MEtatool): semanticweb.cs.vu.nl/amalgame/

Google Refine: http://code.google.com/p/google-refine/

Glossarium CEST: www.projectcest.be

Inventarisatie Terminologiebronnen, DEN (Digitaal Erfgoed Nederland) 2010, http://www.den.nl/terminologiebronnen

Mondeca – Linked Open Vocabularies: http://lov.okfn.org/dataset/lov/

Multilingual websites and multilingual thesauri, Minerva survey 2004-2005, http://www.mek.oszk.hu/minerva/survey/

W3C- Library Linked Data Incubated Group-report:
http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

W3C- Linking Open Data Community Project:
http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

FOAF (Friend of A Friend): http://www.foaf-project.org/

10 APPENDIX 2: DEFINITION OF TERMS AND ABBREVIATIONS

ANSI: American National Standards Institute. The institute oversees the creation, promulgation and use of thousands of norms and guidelines in many different sectors. They have published various standards with guidelines on construction and management of thesauri or controlled vocabularies.

CMS: Collection Management System. The system used by organizations to manage their collection

CC: Creative Commons licences. Creative Commons develops, supports, and stewards legal and technical infrastructure that maximizes digital creativity, sharing, and innovation.

Concept: In the SKOS-model, developed by the World Wide Web Consortium (W3C), terms in a thesaurus are considered *concepts*. This is because, in SKOS, not the term is important, but its hierarchical, equivalent and associative relations, as well as all the additional information it contains, expressed in URIs. A term refers to the lexical string of syllables and vowels, whereas a concept refers to a *unit of thought* expressed in a formal computer language. Because of the formal characteristics of concepts, language barriers can be overcome when linking and retrieving resources.

CSV-file: Comma Seperated-Values. File which stores data in tables in plain text, such as an Excel file.

GLAM sector: sector of galleries, libraries, archives and museums

Interoperability: Interoperability is the ability of systems and software to exchange information. Interoperability can be achieved by following standardized procedures, e.g. by using standards written by the International Organization for Standardization (ISO) for the development of thesauri. When organizations use the same set of rules for a certain activity, they can inter-operate or work together more efficiently, e.g. for creating mutual information systems such as online catalogues.

ISO: ISO-norms are created by the *International Organization for Standardization*. The organization has published more than nineteen thousand international standards covering all aspects of technology and business. The standards are developed by topic, such as *information and documentation*. They are written and supervised by a committee of experts and offer internationally acclaimed rules and procedures. The *ISO 25964-1:2011 (part 1)* for example, contains valuable information on thesauri and interoperability with other vocabularies.

KOS: Knowledge Organization System. A term coined by the Networked Knowledge Organization Systems Working Group. Name for controlled vocabularies in information systems.

LIDO: Lightweight Information Describing Objects. An explicit format to deliver (museum's) object information in a standardized way. It is the result of a collaborative effort of international stakeholders in the museum sector to create a common solution for contributing cultural heritage content to web applications.

Linked Data: Linked data or linked open data (LOD) is a network of information (or digital objects) on the world wide web. This network of digital objects can be obtained when documents, images, thesaurus concepts etc. are represented by URIs. When publishing data in URIs, the data can be shared and reused on the web. Computer systems can easily make links between different resources. The goal of linked open data is to optimize accessible information on the web. URIs can be linked using RDF, a computer language developed by the World Wide Web-consortium (W3C). The basic principles of LOD were coined by Tim Berners-Lee (computer scientist and "inventor" of the world wide web):

1. Use URIs to denote things.
2. Use HTTP URIs so that these things can be referred to and looked up by people.
3. Provide useful information about the thing when its URI is dereferenced, leveraging standards such as RDF, SPARQL.
4. Include links to other related things (using their URIs) when publishing data on the Web.

Mapping: A procedure where elements in a structured dataset (e.g. in a metadata scheme) are linked to elements in another dataset.

Query: A query is a search in a search engine of a local database, online catalogue, web browser etc.

RDF: RDF is short for Resource Description Framework, a format developed by the World Wide Web Consortium (W3C) for exchanging data on the web. It is based on the principle of object and subject and the relation between them. The relation is a predicate. Object, subject and predicate are RDF triples. SKOS, developed to express knowledge information systems such as controlled vocabularies and exchangeable in RDF, is completely built on triples. If guitar is a narrower concept of the concept musical instruments, this will be translated in RDF as: guitar (=subject) → narrower as (=predicate) → musical instrument (=object). Guitar and musical instruments are concepts expressed in URIs. The predicate is expressed by a SKOS property, in this case skos:narrower.

SKOS: SKOS or Simple Knowledge Organization System is a formal data model developed by W3C to enhance linked open data in the (semantic) web. It is a standard that translates knowledge information systems such as thesauri, classification systems etc. in RDF-triples (SKOS/RDF). Controlled vocabularies structure information via hierarchical, equivalence and associative relations and contain scope notes, translations and other additional information on specific terms. This information can be made accessible on the web when the controlled vocabulary is converted to SKOS. In SKOS the term and all the information it contains is expressed in URIs. This is why in SKOS they are called *concepts*. In a controlled vocabulary, the term is important, whereas in SKOS, the URI is important. URIs form the basis of linked data on the web.

Conversion to SKOS requires some technical knowledge of RDF and SKOS. This is why the Linked Heritage and AthenaPlus projects developed a Terminology Management Platform (TMP), an open-source tool where controlled vocabularies can be imported and link them to other resources using SKOS.

An introduction to SKOS by Regine Stein from the Philipps-Universität Marburg
- Bildarchiv Foto Marburg - is available on slide-share: http://www.slideshare.net/ISOCIL/skos-handson-workshop-tutorial-by-regine-stein

Semantic Web: The Semantic Web is a collaborative movement led by the international standards body, the World Wide Web Consortium (W3C). The semantic web is built on the principle of sharing and reusing data on the web to achieve better search results, irrespective of language. This can be done by automatically linking “separate” data on the web. When every concept is interlinked as an equivalent, synonym, broader or narrower concept or via any other relator, the web can optimize search results. This will engage greater visibility and easier access to information.

Terminology: general name for controlled vocabularies. Each descriptor is a preferred term with an unambiguous, non-redundant definition. Descriptors in a controlled vocabulary can have hierarchical, equivalent or associative relations. A controlled vocabulary is managed by an authority, this can be a thesaurus manager or a centralized organization responsible for managing the vocabulary. Controlled vocabularies allow a standardized way of indexing collections in a local database or online catalogue. It is also a powerful tool for web search queries and for sharing data on the web. Thesauri, classification systems, taxonomies and subject headings are types of controlled vocabularies. They are also referred to as authority lists.

Thesaurus: A thesaurus is a type of controlled vocabulary. It is considered the most elaborate form of vocabulary, as it contains a large amount of information. Terms in a thesauri are related to each other by hierarchical, equivalent and/or associative relations. A hierarchical relation means that one term is considered broader or narrower than another, expressing for example a “sort of” relation: a *guitar* is considered a narrower term of a *musical instrument* because a guitar is a “sort of” musical instrument. It is a vertical relation. An equivalent relation means that several terms are considered equal, but one term is to be preferred to another. For example, *house* and *dwelling* are synonyms, but in a thesaurus one term will be preferred and the other will be alternative. This relation is horizontal. An associative relation represents non-direct relations: the term is not a narrower nor a broader term, nor is it a synonym, but there is a relation anyhow. *Guitar* can be a narrower term of *musical instruments* and *guitar tabs* can be

a narrower term of *sheet music*. Even though they do not have a parent-child relationship, *guitar* can be linked to *guitar tabs* via an associative relation.

Terms in a thesaurus are considered unique and can have a unique identification number (reused in a URI). Their meaning and use are described in scope notes.

TMP: Terminology Management Platform, an open-source tool developed to create new or import existing terminologies, map them with other terminologies using SKOS properties and export them in RDF. A first version of the tool was launched in the Linked Heritage project, a production version will be made available in the AthenaPlus project.

URI: Uniform Resource Identifiers are references to digital objects. These objects can be images, texts, movies, but also metadata-records in a collection management system. There are two types of URIs. A URL (Uniform Resource Locator) is an identifier of the place where something is located and a URN (Uniform Resource Name) give the record a fixed name[32]. The URI's should also be persistent identifiers.

W3C: World Wide Web Consortium. An international community where Member organizations, a full-time staff, and the public work together to develop Web standards. Led by Web inventor Tim Berners-Lee and CEO Jeffrey Jaffe, W3C's mission is to lead the Web to its full potential.

XML: XML or Extensible Markup Language is a computer language standard developed by W3C that defines a set of rules for encoding documents in a format that is human-readable and machine-readable. RDF/XML is an application of XML, created to express RDF as an XML-document.

[32] Glossarium CEST: www.projectcest.be