

ECP-2005-CULT-038099

ATHENA

Guidelines for mapping into SKOS, dealing with translations

Deliverable number	<i>D4.2</i>
Dissemination level	<i>Public</i>
Delivery date	<i>30th July 2010</i>
Status	<i>In progress</i>
Author(s)	<i>Marie-Véronique Leroi (MCC/DREST), Johann Holland (Michael Culture Aisbl)</i>



eContentplus

This project is funded under the eContentplus programme¹, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

Table of Contents

1. EXECUTIVE SUMMARY	3
2. INTRODUCTION	4
2.1. CONTEXT AND OBJECTIVES	4
2.2. AUDIENCE AND REQUIRED SKILLS	6
2.3. BASIC CONCEPTS	6
3. SKOS AND TERMINOLOGY MAPPING	12
3.1. SKOS	12
3.2. METHODOLOGY FOR MAPPING	18
4. WP4 EXPERIMENT	20
4.1. PROCESS AND ISSUES	21
4.2. BENCHMARK RESULTS	24
4.3. MEANS	25
4.4. ATHENA FORMAT	27
4.5. EXPERIMENT RESULTS	31
5. GUIDELINES	38
5.1. BENEFITS IN USING SKOS	38
5.2. GUIDELINES FOR SKOSIFICATION	39
5.3. MAPPING	45
6. CONCLUSIONS	49
6.1. WHAT WOULD BE AN IDEAL TOOL?	49
6.2. PERSPECTIVE	50
7. ANNEXES	51
7.1. ACRONYMS	51
7.2. PROCESS AND ISSUES	51
7.3. BENCHMARK: TOOLS	62
7.4. POOL PARTY SKOS VALIDATOR	66
7.5. SCHEMATIC VIEW OF THE MICHAEL SUBJECT LIST FOR ARCHITECTURE DOMAIN	67
7.6. SCHEMATIC VIEW OF THE RMCA THESAURUS FOR ARCHITECTURE DOMAIN	67
7.7. SCHEMATIC VIEW OF THE PICO THESAURUS FOR ARCHITECTURE DOMAIN	68
7.8. SCHEMATIC VIEW OF THE ATHENA THESAURUS	69

1. Executive summary

This deliverable is part of ATHENA Workpackage 4 (WP4) and addresses European museums. It aims to present a set of guidelines for the mapping of terminology resources into SKOS (SKOSification) dealing with translations (multilingualism). The deliverable is structured as follows:

- *Executive Summary*: A short summary of the deliverable
- *Introduction*: Explaining the context of the whole work package in which the deliverable stands, the objectives of the task that the deliverable relates to, the audience for these guidelines and the skills needed to apply them, and the basic concepts the reader should have a grasp of in order to have a good understanding of the deliverable content.
- *SKOS and terminology mapping*: Presenting SKOS features, the ATHENA Format, and methodology for mapping.
- *Experiment within the WP4*: Presenting, in detail, the aims and objectives of this experiment, and the methodology that was set up for the elaboration of the first core of the ATHENA Thesaurus.
- *Guidelines*: For European museums in relationship to SKOSification and mapping of multilingual terminologies.
- *Conclusions and perspectives*: The main results and a proposal of work for the future.

2. Introduction

2.1. Context and objectives

2.1.1. WP4 short introduction

According to the ATHENA *Description of Work* WP4 should explore the current practices in the field of terminology adopted by European museums, to be compared with those used in other sectors of cultural heritage and in cross-domain portals, in order to guarantee semantic interoperability within Europeana. It is working on multilinguality issues by surveying and integrating existing multilingual tools and ensuring the alignment between the museums' terminologies and the ATHENA SKOS thesaurus.

2.1.2. Task introduction

We already have an overview of terminology use in European museums¹ and have created a first set of recommendations based on an analysis of this overview².

In order to enhance these recommendations we launched an experiment the first phase of which is now complete. The experiment consists in building from different existing terminology resources a common thesaurus that is the Athena Thesaurus. This process is achieved on the basis of specific criteria and allows us to test our recommendations and guidelines. A summary of the results was presented in a workshop that took place in Paris on the 25th June 2010. This deliverable presents more in depth all those results. It will be helpful for the follow-up of the WP4 and will ensure the consistency of the final recommendations for the SKOSification of the terminology resources in order to allow semantic interoperability with Europeana. By SKOSification we mean the transformation of a terminology resource into SKOS.

Indeed Europeana in its data ingestion process is also gathering the corresponding terminology resources. In order to provide the semantic data layer to its contents, Europeana requires these terminology resources to be provided as Linked Open Data (LOD) links or as SKOSified versions. As mentioned in the White paper from Europeana³, these requirements will help to propose to the end-users knowledge, e.g. information in context rather than "simple" data.

¹ ATHENA Deliverable D4.1

- PDF version: <http://www.athenaeurope.org/getFile.php?id=398>
- Wiki: <http://www.athenaeurope.org/athenawiki/> (section D4.1: resources)

² <http://www.athenaeurope.org/athenawiki/index.php/Recommendations>

³ http://version1.europeana.eu/c/document_library/get_file?uuid=cb417911-1ee0-473b-8840-bd7c6e9c93ae&groupId=10602

2.1.3. Relationship with the work of other ATHENA Work Packages

Relationship with LIDO

Our work takes into account LIDO as data model for ATHENA. LIDO was developed in WP3 and in collaboration with WP7. LIDO has been designed in order to take into account every semantically enriched vocabulary notably SKOS then the interoperability between the LIDO metadata scheme and the ATHENA format for terminology. WP3 is taking part to the WP4 working group and can make sure that there is no contradiction between the two models.

LIDO¹ stands for Lightweight Information Describing Objects and is the result of a joint effort between existing initiatives: CDWA Lite, Museumdat² and SPECTRUM³. The first point to make about LIDO is that it is a XML harvesting schema. It should not be used as a basis for a collection management system. It is for delivering metadata for use in the service environment of an organisation's online collections database, portals, and aggregations, including Europeana itself. In particular it does not support such activities as loans and acquisition. Its strength lies with its ability to support the full range of descriptive information about museum objects.

LIDO is made up of a nested set of 'wrapper' and 'set' elements which structure records in culturally significant ways. An important part in its design is the concept of events taken from the CIDOC CRM. So, for example, the creation, collection, and use of an object are defined as events which have associated entities such as date, places and actors. These are all represented in a consistent way in the schema.

The structural elements of LIDO contain 'data elements' which hold the information that is being harvested and ultimately delivered to the user of the service environment.

LIDO also allows for the recording of information about the sources for data (e.g. in a book) and controlled terminology (e.g. the identification code for a term in a thesaurus).

Relationship with ATHENA Ingestor Server

Our work also takes into account the ATHENA Ingestion Server⁴, a web service developed by WP7 that implements LIDO as data model. WP7 is also taking part to the WP4 working group and is a support for evaluating the possibility of an Ingestor-like tool for the management of terminology.

With this Ingestor users can:

1. Map their metadata to LIDO.

¹ See D3.3 Lido-0-8.pdf at: <http://www.athenaeurope.org/getFile.php?id=535>

² http://museum.zib.de/museumdat/cdwalite_and_museumdat.pdf

³ <http://www.collectionstrust.org.uk/schema>

⁴ More information on the ATHENA Ingestor on the training section of the ATHENA website: <http://www.athenaeurope.org/index.php?en/159/training>

2. Declare their collections are available to be ingested into Europeana.

One feature of the Ingestor is that it can take into account the organisational structure of a cultural institution, and therefore make it possible for users to be managed efficiently. It also has a preview of how records will look in Europeana.

2.2. Audience and required skills

The audience for this deliverable are the staff in European museums. It aims to provide a set of guidelines for the SKOSification and mapping of terminologies. This type of work is usually carried out by those with skills in information engineering, specifically in computing and knowledge management. Most museum staff do not have these skills, and the cost to get them is high.

To overcome this barrier we will introduce the issues, methods and tools that the reader needs to be aware of in order able to take full advantage of the guidelines. They will also define the areas of work that might require specialist technical help. The point is also to define the tasks and work that might require a technical help and make non-experts aware of these aspects.

2.3. Basic concepts

2.3.1. About terminology

In the first deliverable of the WP4 we identified different types of terminology resources. Here is a reminder:

So far we have used the word “terminology” for the resources used by the museums in describing their collections. However “terminology” might be ambiguous. Strictly speaking “terminology” is a discipline for the studying of terms and their use within a specific domain; but a “terminology” can also refer to the resource resulting from this discipline. However “terminology” is still the most common word used for the different types of resources:

- Lexicon
- Dictionary
- Folksonomy
- Glossary
- Classification
- Taxonomy
- Thesaurus
- Controlled vocabulary
- Terminology
- Ontology

The type of resource used is highly connected to its purpose. An information retrieval tool and a knowledge management tool may not use the same kind of resource:

- Some of the resources mentioned above (e.g. lexicons and dictionaries) are mainly dedicated to linguistic concerns, not for a specific domain, and for the use of human beings only. Lexicons and dictionaries deal with words and not with terms.
- Some other resources such as folksonomy are directly managed by non-expert users in order to improve access to the information in a collaborative way.
- The other resources mentioned (e.g. classification, thesaurus, ontology) are more formal, being presented as alphabetical lists or networks of terms, and they can be specific to a domain. They can be used by computational programs for different purposes such as indexing or translating but are also meant to be handled by experts of a domain. Most of these resources deal with terms or concepts rather than words.

Here we focus on the “thesaurus” as this kind of resource was recommended in the D4.1. Indeed thesauri can be easily used for the mapping of in-house terminologies to a reference one. Thus here we provide a definition to make it explicit in which sense we use the term.

Thesaurus :

A thesaurus can be defined as “a networked collection of controlled vocabulary terms”. Thesauri allow the connection of terms using several types of relationships which can be hierarchical, associative, equivalence or definition. This means that a thesaurus uses associative relationships in addition to parent-child relationships. A parent-child relationship is expressed by a Broader Term (BT) /Narrower Term (NT) feature. Associative relationships in a thesaurus such as “Related Term” (RT) (e.g. term A is related to term B) are used to express relationships that are neither hierarchical nor equivalent. Equivalence is expressed by the USE (e.g. preferred term)/ Used For (UF) (e.g. non-preferred term). Additional information such as definition or remark can be included in a Scope Note (SN). The equivalence relationship is especially useful within multilingual thesauri.

Several standards have been established to provide guidance for the elaboration of this kind of terminology. The standards are:

- *ISO 2788:1986: Guidelines for the establishment and development of monolingual thesauri:* This standard recognized by the International Organization for Standardization) consists of recommendations for the establishment and development of consistent indexing practice within an organization or a consortium. The standard assumes that indexing is being done by humans using natural language to select indexing terms. It is most suitable for cataloguing and descriptive metadata. The standard only deals with monolingual thesauri and is based on the use of preferred terms or indexing terms and non-preferred terms or synonyms
- *ISO 5964: 1985: Guidelines for the establishment and development of multilingual thesauri:* This ISO standard extends the scope of ISO 2788 to cover particular considerations for multilingual thesauri development for the establishment of consistent indexing practice within an organization or consortium. Like ISO 2788, the standard assumes that indexing is being done by humans using normal language, and is based on the concept of preferred terms or indexing terms and non-preferred terms or synonyms. The standard covers general problems, language problems and management decisions required when establishing a multilingual thesaurus. It considers the issues of vocabulary

control, establishing equivalent terms across different languages, relationship between terms, display of terms and relationships, form and contents and organization of work.

- *ANSI/NISO Z39.19-2003: Guidelines for the Construction, Format, and Management of Monolingual Thesauri*: This Standard presents guidelines and conventions for the contents, display, construction, testing, maintenance, and management of thesauri. It covers all aspects of constructing thesauri including extensive rules and guidelines for term selection and format, the use of compound terms, and establishing and displaying various types of relationships among terms. This standard focuses on monolingual thesauri; it has been revised in 2005 in order to extend its scope to controlled vocabularies e.g. lists of controlled terms, taxonomies, thesauri.
- *BS8723: Structured Vocabularies for Information Retrieval*: This standard, which is a British adaption of the ISO 2788, intends to take into account every kind of terminology not only thesauri and focuses also on the interoperability between vocabularies.

2.3.2. Semantic Web

The Semantic Web (part of Web 3.0) is “the Web of data with meaning in the sense that a computer program can learn enough about what the data means to process it”¹. It provides “a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by World Wide Web Consortium (W3C) with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming. It was proposed by World Wide Web inventor Tim Berners-Lee”².

As we can read on Wikipedia³:

“Semantic Web is a term coined by World Wide Web Consortium (W3C) director Tim Berners-Lee. It describes methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web.”

The availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user.

While the term “Semantic Web” is not formally defined it is mainly used to describe the model and technologies proposed by the W3C. These technologies include the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

¹ <http://www.w3.org/People/Berners-Lee/Weaving/glossary.html>

² <http://www.uen.org/core/edtech/glossary.shtml#S>

³ http://en.wikipedia.org/wiki/Semantic_Web

Many of the technologies proposed by the W3C already exist and are used in various projects. The Semantic Web as a global vision, however, has remained largely unrealized and its critics have questioned the feasibility of the approach. These critics mainly stem from ethical issues (respect of privacy) and practical feasibility (general user behavior and personal preferences).

In addition other technologies with similar goals, such as microformats, have evolved, which are not always described as “Semantic Web”.

In order to overcome these critics and apprehension the Semantic Web provide a set of standards and tools which enable a machine to process knowledge itself instead of text using processes similar to human reasoning and inference for obtaining more meaningful results.

2.3.3. Linked Data

As a first definition we can say¹:

“In Semantic Web terminology, Linked Data is the term used to describe a method of exposing and connecting data on the Web from different sources. Currently, the Web uses hypertext links that allow people to move from one document to another. The idea behind Linked Data is that hyperdata links will let people or machines find related data on the Web that was not previously linked. The main point is that the focus is more about data and how to create and maintain links between these data than documents and links between documents.”

Here is a more “official” definition from Tim Berners-Lee²:

“The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the Web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the Web grow:

- 1. Use URIs to identify things (anything, concrete or abstract things, not just documents)*
- 2. Use HTTP URIs so that people can look up those things.*
- 3. Provide useful information using standards (RDF*, SPARQL) when someone looks up a URI*
- 4. Include links to other URIs (RDF links generally) to enable the discovery of related information.”*

¹ http://www.webopedia.com/TERM/L/Linked_Data.html

² <http://www.w3.org/DesignIssues/LinkedData.html>

2.3.4. Formats

In order to be part of the Linked Data ‘cloud’ and use Semantic Web technologies the terminology of an institution has to be in compliant format. When you want to represent or model your terminology, and to exploit it on the Web, you have to use a format standard. The most commonly used format standards are SKOS, OWL, RDF, and XML. Some of them can be combined, and some of them can be wrapped by others. Using a format standard will result in the metadata, expressed with your terminology, being effectively represented in a way the Web technologies can recognize and interpret.

Below are brief descriptions of these format standards with the aim of a better understanding of their connections.

XML¹

XML (Extensible Markup Language) is a set of rules for encoding documents in machine-readable form. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all free to use open standards.

XML's design goals emphasize simplicity, generality, and usability over the Internet. It is a textual data format, with strong support via Unicode for the languages and scripts of the world. Although XML's design focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

There are many programming interfaces that software developers may use to access XML data, and several schema systems designed to aid in the definition of XML-based languages.

RDF²

The Resource Description Framework (RDF) is a family of W3C specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources, using a variety of syntax formats.

The RDF data model is based upon the idea of making statements about resources (in particular Web resources) in the form of triples. Triples are the expressions of statements about resources which are presented as subject-predicate-object expressions. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

The RDF specification is based on the XML encoding.

¹ <http://en.wikipedia.org/wiki/XML>

² http://en.wikipedia.org/wiki/Resource_Description_Framework

OWL¹

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. The languages are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium and has attracted academic, medical and commercial interest.

In October 2007, a new W3C working group was started to extend OWL with several new features as proposed in the OWL 1.1 member submission. This new version, called OWL 2, soon found its way into semantic editors such as Protégé and semantic reasoners such as Pellet, RacerPro and FaCT++. W3C announced the new version on 27 October 2009.

The OWL family contains many species, serializations, syntaxes and specifications with similar names. This may be confusing unless a consistent approach is adopted. OWL and OWL2 will be used to refer to the 2004 and 2009 specifications, respectively. Full species names will be used, including specification version (for example, OWL2 EL). When referring more generally, OWL Family will be used.

OWL is based on the RDF specification.

SKOS

In this set of formats, SKOS is more and more required by web services. Europeana for instance has decided to format in SKOS all the metadata they harvest for a homogeneous and effective exploitation of the resources, of the data and their related descriptions. SKOS is based on the RDF specification and enable a migration towards OWL ontologies.

SKOS is not a formal knowledge representation language since literally a formal knowledge is expressed as sets of axioms and facts which are the main features of a formal ontology. SKOS is rather used for modeling controlled vocabularies such as thesauri or classifications which are of a different nature than ontologies. The ideas or meanings described by thesauri or other kinds of terminology are referred to as “concepts” even if from the ontological point of view a concept is defined in a different way.

The next section defines more precisely what SKOS is and what its features are.

¹ http://en.wikipedia.org/wiki/Web_Ontology_Language

3. SKOS and terminology mapping

3.1.SKOS

Europeana requires that the object and collection descriptions of the museums are expressed with a terminology resource designed or converted in the SKOS format. Thus we document here what SKOS is. Then we introduce a specific format, ATHENA Format, that is currently set up by the WP4. This ATHENA Format is detailed as a SKOS-compliant reference format for all the museums' terminologies.

SKOS stands for Simple Knowledge Organization System and is the result of several years of work in the field of the Semantic Web. SKOS was first designed within the Semantic Web Advances Development for Europe Project (SWAD-Europe) working group before being submitted to the W3C¹. SKOS has been acknowledged as a W3C recommendation in August 2009.

The official documents presenting the SKOS data model² often make an opposition and parallelism between unstructured data/human-readable and structured data/machine-readable data. The aim of SKOS is the better organization of unstructured data and the making of it meaningful.

The overall aim of SKOS, in conformity with the spirit of the Semantic Web, is to bring together information from different fields and communities of practice. The purpose of SKOS is to share and link knowledge organization systems (KOS) via the Web and allow semantic interoperability between terminologies of different types and languages.

3.1.1. Main features

As a Semantic Web compliant format, SKOS is concept-oriented. This means that the fundamental element of a terminology designed in SKOS is the concept and not the term that expresses this concept.

The SKOS data model consists of a basic structure that can be extended by specific classes for detailing lexical parts or semantic relations between the concepts of the terminology.

The SKOS reference publication summarizes the main features of the SKOS model as follows³:

¹ World Wide Web Consortium: International community that develops standards to ensure the long-term growth of the Web. <http://www.w3.org/>

² SKOS Reference : <http://www.w3.org/TR/2009/REC-skos-reference-20090818/> (18th of August 2009)

SKOS Primer : <http://www.w3.org/TR/skos-primer/> (18th of August 2009)

³ SKOS Reference : <http://www.w3.org/TR/2009/REC-skos-reference-20090818/> (18th of August 2009)

“Using SKOS, can be identified using URIs, with lexical strings in one or more natural languages, assigned (lexical codes), with various types of note, and organized into informal hierarchies and association networks, aggregated into, grouped into, labeled and/or ordered , and to concepts in other schemes.”

SKOS data are expressed as RDF triples. This means that concepts may be subject or object and related via a SKOS property which would be the predicate.

As RDF triples, SKOS concepts can be identified using URIs. These URIs can be defined according to standard persistent identifier systems. The SKOS data model doesn't require the use of persistent identifiers but in a Linked Open Data perspective, their use is highly recommended. Persistent identifiers will be described more precisely in the following sections.

The SKOS data model consists in three main components: classes, properties and relations. These three components always start with the prefix “*skos:*”. The distinction between a class and a property is done through the case: the element following the “*skos:*” prefix starts with an upper-case character when it is a class, e.g. *skos:Concept* and *skos:ConceptScheme* are classes; if the element following the “*skos:*” prefix starts with a lower case character, this means that the element is a property and not a class. For example *skos:prefLabel* is a property.

Concept

SKOS is a concept-oriented data model therefore the concept is the central element of the terminology. From a terminology point of view a concept can be defined as an idea, notion or unit of thought. A concept in SKOS is introduced as a class *skos:Concept*. Some bridges between the SKOS data model and the OWL one are available for a better interoperability. The *skos:Concept* class is an instance of *owl:class* which is a class from the OWL data model so that connections between the two data models are enabled.

SKOS concepts can be brought together into two classes:

- SKOS concept scheme
- SKOS collections

Concept Schemes

A concept scheme is a way to bring together several concepts. A concept scheme is introduced by the class *skos:ConceptScheme*. An individual concept scheme roughly corresponds to the notion of an individual thesaurus, classification scheme or any other knowledge organization system.

It is important to mention that a same concept can be part of more than one concept scheme.

Concept collections

A collection is a group of SKOS concepts. A collection is introduced by the main class *skos:Collection*. Although another class *skos:OrderedCollection* can also be used in the case where the order of the concepts within the collection has an importance.

The notion of collection is different from the concept scheme. For the migration of a thesaurus for example, the whole could be considered as a concept scheme where several thematic groups of concepts could be designed as collections.

Identifiers

Each concept must be identified in a unique way in order to avoid any ambiguity. As it is the case in the RDF language and as a general principle of the Semantic Web and Linked data, it is recommended to use HTTP URIs in order to identify the concepts of terminology.

The identifiers are introduced by a specific RDF property *rdf:resource* which is used each time that a new concept is introduced or semantic relations or mapping to other concepts are included in the description of the concept.

Labels

The SKOS model focuses on concepts therefore there is a distinction between the concept itself and the terms that may be used to express this concept. Terms referring to a concept can be expressed via lexical labels according to the SKOS data model. A lexical label is a string of Unicode characters which allows you to have a term in any language with or without Latin characters.

The SKOS data model defines 3 types of lexical label:

- Preferred label
- Alternative label
- Hidden label

The use of these different types of label enables the understanding of the concept and is useful for human-readable knowledge representation. The use of labels is not mandatory in the SKOS datamodel but is highly recommended especially for maintenance purposes.

Preferred label

The preferred label, introduced in the SKOS data model as the *skos:prefLabel* property, corresponds to the notion of descriptor from the standards for the elaboration of thesauri (ISO 2788 and ISO 5694).

The SKOS data model does not allow there to be more than one preferred label in the same language.

Alternative label

Alternative labels, introduced as *skos:altLabel* property, are mainly used to give synonyms to the preferred label or other ways to refer to this preferred label, e.g. different spellings or acronyms.

The SKOS model does not forbid the exclusive use of alternative labels instead of one preferred label and many alternative labels.

Hidden label

Hidden labels, introduced by the *skos:hiddenLabel* property, may be used for mentioning the misspellings of preferred or alternative labels but also for mentioning obsolete forms of a term.

Alternative and hidden labels correspond roughly to the USE and UF (Used For) indicators defined in the ISO standards for thesauri.

By definition, hidden labels are not visible but are very useful for the retrieval.

Obviously the SKOS data model does not allow the use of the same string of characters as a preferred, alternative or hidden label in the same language.

An extension to the SKOS model, SKOS-XL, is proposed for modeling more precisely the labels and including morphologic or syntactic information on labels.

Notation

Another property is available for expressing notations which are different from labels. Notations are symbols or codes that are not recognizable or understandable in any natural language. Notations are different from labels which usually are words or expressions understandable in any natural language. The *skos:notation* can then be used for example in the case of classifications where a code refers to a term referring itself to a concept. The notation can be more convenient than using an alternative label since it is considered as unambiguous and language independent.

Documentation properties

The SKOS model offers a variety of possibilities to provide information related to concepts. Different types of notes can be used to give the most accurate information. These notes can be of different natures (plain text, image, quotes ...) and be used without any restriction.

The different types of notes that can be used to document a concept are:

- Note (*skos:note*)
- Change note (*skos:changeNote*)
- Definition (*skos:definition*)
- Editorial note (*skos:editorialNote*)
- Example (*skos:example*)
- History note (*skos:historyNote*)
- Scope note (*skos:scopeNote*)

The *skos:note* can be used to provide general documentation on a concept. All the other types are specializations of this general property.

The *skos:changeNote* and *editorialNote* are mainly useful for the purpose of administration and maintenance. The *skos:definition*, *skos:example*, *skos:historyNote* are useful for providing information on the concept for a better understanding of its meaning.

As for labels, documentation properties can be provided in different languages by using language tags with the *xml:lang* attribute.

Semantic relations

The power of the SKOS model lies in the semantic relations that can be used to connect between different concepts. These semantic relations play a crucial role for defining concepts. There are two different categories of semantic relation:

- Hierarchical
- Associative

Hierarchical relations

Hierarchical relations are introduced via two properties, *skos:broader* and *skos:narrower*. The *skos:broader* property is used to assert that a concept has more general meaning. *skos:narrower* is the inverse property used to assert that a concept has a more specific meaning.

One concept can have more than one broader concept or more than one narrower concept.

It is important to note that these two properties only assert direct/immediate hierarchical link between two concepts. In order to enable non-immediate link between two concepts, the SKOS model provides two other properties that are transitive. The graph below provides an example of this case:

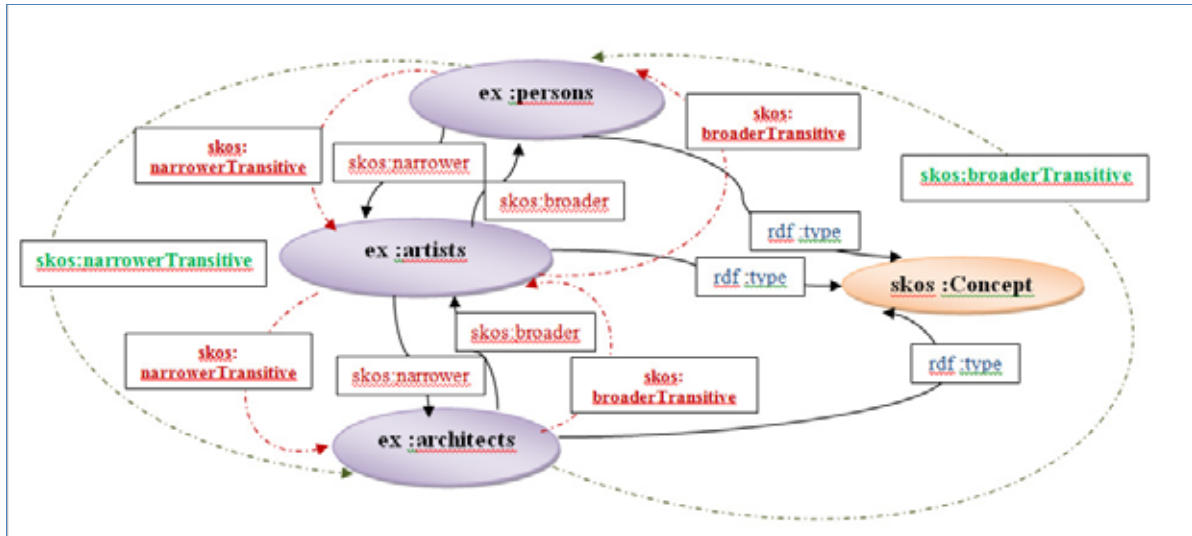


Figure 1: Graph illustrating the *skos:NarrowerTransitive* property

The use of *skos:broaderTransitive* property is necessary to assert that “persons” is broader than “architects”.

As for the *skos:broader* and *skos:narrower*, the properties *skos:broaderTransitive* and *skos:narrowerTransitive* are the inverse of each other.

Associative relations

The property *skos:related* is used to assert an associative link between two concepts. This property may be useful to make a link between a concept and another one which is neither an equivalent nor a broader/narrower concept. It is important to note that the *skos:related* property is symmetric.

For example it is possible to state that “monuments” is related to “buildings”. Then by symmetry it is also possible to state that “buildings” is related to “monuments”.

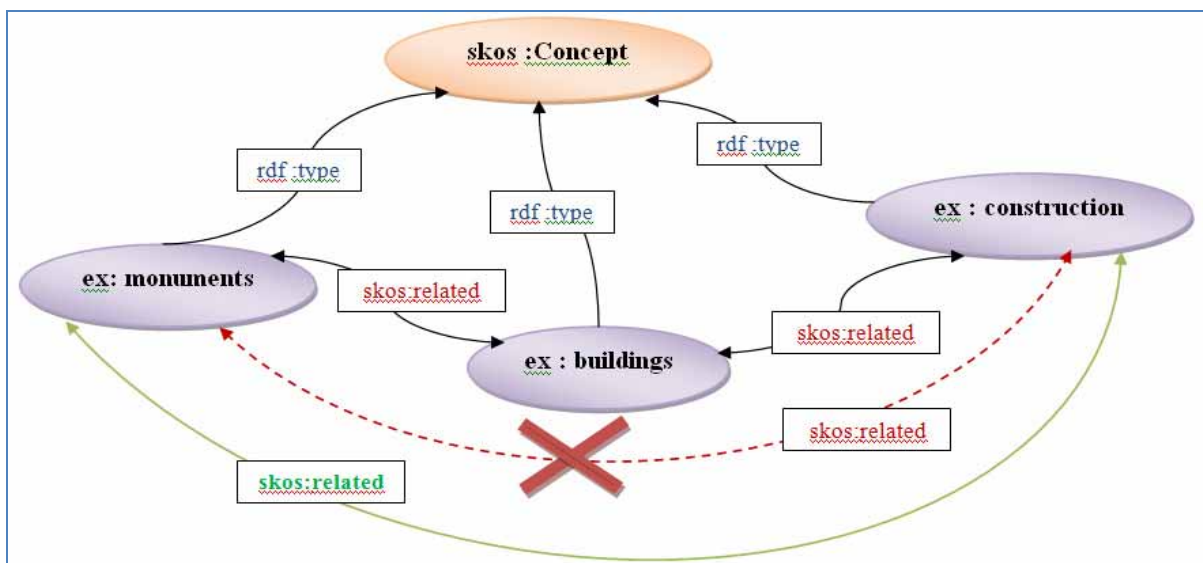


Figure 2: Graph illustrating the *skos:related* property

skos:related is not a transitive property. This means that referring to the previous example, if we state that “buildings” is related to “construction”, it does not follow that “monuments” is related to “construction”. The dotted arrow illustrate that this semantic relation between “monuments” and “construction” cannot be deduced via the transitivity but have to be asserted separately.

It is very important to keep in mind that, according to the guidelines provided in ISO 2788 and BS8723, mixing associative relations and hierarchical relations is not consistent with the SKOS data model. Therefore a special attention must be paid to the semantic relationships between concepts.

Mapping

The SKOS data model provides several mapping properties for making alignment between concepts from different concept schemes. These properties are :

- *skos:closeMatch*
- *skos:exactMatch*
- *skos:broadMatch*
- *skos:narrowMatch*
- *skos:relatedMatch*

As for semantic relations between concepts, the mapping properties can be associative or hierarchical. The *skos:broadMatch* and *skos:narrowMatch* properties are used for a hierarchical mapping link between concepts whereas the *skos:relatedMatch* property is used for an associative one. Exactly as for semantic relations, *skos:broadMatch* is the inverse property of *skos:narrowMatch*.

The properties *skos:closeMatch* and *skos:exactMatch* are used to make a mapping link between concepts that are very similar or equal so they can be used interchangeably. The *skos:exactMatch* property is transitive and symmetric.

Mapping properties are used rather than semantic relations in order to make mapping links between concepts from different concept schemes. In the case of a same concept scheme semantic relationships will be used instead of mapping properties.

As for semantic relations, there may be some conflicts in mixing hierarchical mapping properties with associative ones.

3.2. Methodology for mapping

3.2.1. Mapping process

With regards to the context we have described above, we call “mapping process” the way to relate different terminologies together. The term “mapping” is usually used to refer to the alignment of data models or metadata schemes which are, in other words, a grammar. In the context of this work package, we only focus on the mapping of terminology which is the

semantic part of this grammar. The process and methodology for mapping are the same in both cases but the purpose and the use cases may be different.

In the case we are concerned with, namely terminology mapping, the objective is to relate in-house terminologies with a reference thesaurus compliant with Europeana requirements. Basically the mapping process consists in the alignment of terms between the terminologies. Such mapping is the result of a multiplicity of particular relations. Every relation between one term of an in-house terminology and one term of the reference thesaurus may have a strict type among all the possibilities given by the format for alignment. However every relation may have a specific purpose out of a systematic organisation. The reason or motivation of every relation may be made explicit, or not. The alignment may relate terms in different languages so that multilinguality is taken into account without using a specific language for translation.

In the case of mapping of terminology resources we identified several theoretical cases.

3.2.2. Theoretical cases

Here we give an overview of all the theoretical cases of mapping we can envisage in the case of ATHENA project. This overview is the combination of two main sources:

- “About alignment and linking of terminologies”: An article of Jean Delahousse (Mondeca) (article available in French¹)
- “Guidelines for Multilingual Thesauri”: A report of IFLA (International Federation of Library Associations and institutions)²

As proposed in the Mondeca article about mapping, let us distinguish between “alignment” and “linking”. We talk about “alignment” when the two terms mapped together are strictly describing the same domain. We talk about “linking” when these two terms can describe different domains, and there is a reason for their mapping.

Keeping in mind this distinction, we can consider:

- The strict alignment of terms without any necessary precision
 - Because the terminologies describe the same domain
 - And/or the terms are strictly equivalent in the same language
 - And/or the terms are strictly equivalent in different languages
- The linking of terms with a need of precision at the terminology level (through concept schemes or others)
 - Because the terminologies describe complementary domains
 - And/or the terminologies describe the same domain but with more or less precision

¹ <http://mondeca.wordpress.com/2009/06/29/sur-l%E2%80%99alignement-et-la-mise-en-correspondance-de-terminologies/>

² <http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf>

- And/or the terminologies describe the same domain but according to different points of view
- The linking of terms with a need of precision at the terms level (through notes or others)
 - Because the terms are not strictly equivalent in the same language even if the domains are the same
 - Because the terms are not strictly equivalent in the different languages even if the domains are the same

4. WP4 experiment

In the follow-up to the D4.1, we decided to start an experiment about terminology management. One objective of this experiment was to have a better understanding of the possibilities for all the tasks concerned with the construction of a reference thesaurus, the ATHENA Thesaurus. For the management of a terminology resource, we identified six steps described in the next section. Even if SKOSification and mapping are just two of these six steps, this experiment was especially focused on these specific areas. It appeared to us that we should put them into a larger context to provide guidelines that are more precise and valuable.

So the first aim of this experiment consisted of understanding better the use cases and the logical processes concerned by what we could call “terminology management by European Museums in regards with Europeana requirements”. Consequently the objective was to raise all the related issues that museums have to deal with them¹. This effort enabled us to make explicit our understanding of the museums’ needs for tools and functionalities. So we structured a grid of needs which allows us to organise a benchmark² of all the existing initiatives and tools that could be helpful to meet these needs.

We have taken into account the information we got during a technical workshop³ that was organised about the benchmark. For that workshop we asked the speakers to follow our grid of needs as an outline for their presentation. By improving our knowledge on the selection of tools, the main objectives were to obtain an effective overview of how we can, partially or totally, meet user needs with existing technologies.

Described in the following sections are:

- The process of the ATHENA Thesaurus construction, and all the issues we raised
- The methodology we used for creating a first version of the ATHENA Thesaurus
- And the results of the experiment.

¹ http://www.athenaeurope.org/athenawiki/index.php/Process_and_issues

² <http://www.athenaeurope.org/athenawiki/index.php/Benchmark>

³ the workshop took place in Paris at 25th June 2010

4.1. Process and issues

For the construction and maintenance of the ATHENA Thesaurus, we identified six main steps:



1. Registration of an in-house terminology in the platform repository
2. SKOSification of this terminology
3. Search and navigation into a network of lists of terms (i.e. the ATHENA Thesaurus)
4. Mapping of the terminology with the ATHENA Thesaurus
5. Enrichment of the ATHENA Thesaurus
6. Collaborative moderation of updates/modifications of the thesaurus

All the details about what they precisely are, and about the related issues, are given in the Annexes of this deliverable. Below we just give details about the two steps that this deliverable deals with: SKOSification and mapping.

4.1.1. SKOSification use cases

You work for a European museum, and you are in charge of the mapping of the terminology used in your institution with a reference thesaurus compliant with Europeana requirements. Now your terminology has to be SKOSified before any mapping.

First case:

Your in-house terminology is already SKOSified and you want to check if the SKOSification result is correct. To check if it is well-SKOSified, you use a web service enabling you to precisely know what mistakes you have made based on the SKOS model. Then you take into account this feedback and use that web service, or another one, to refine the SKOS version of your terminology. You repeat the process until your terminology is well-SKOSified.

The W3C offers on line a validation tool but it doesn't take into account the latest version of the SKOS model¹. Pool Party, a thesaurus management system, offers online SKOS services² for converting and checking the consistency of your SKOS thesaurus.

Second case:

Your in-house terminology is not SKOSified. To SKOSify your terminology, you use a web service, or carry out the conversion with other technical tools, enabling you to express correctly based on the SKOS data model your terminology features and terms. Then as described above, you check the result in order to verify that your terminology is well-SKOSified, and iteratively refine it while it is necessary.

¹ <http://www.w3.org/2004/02/skos/validation>

² <http://demo.semantic-web.at:8080/SkosServices/index>

Technical remark:

Technically speaking, we met different situations where a specific schema for transformation had to be defined. Indeed when you want to transform one resource into SKOS you can apply schemas. During our work we noticed several cases illustrated by the following schema:

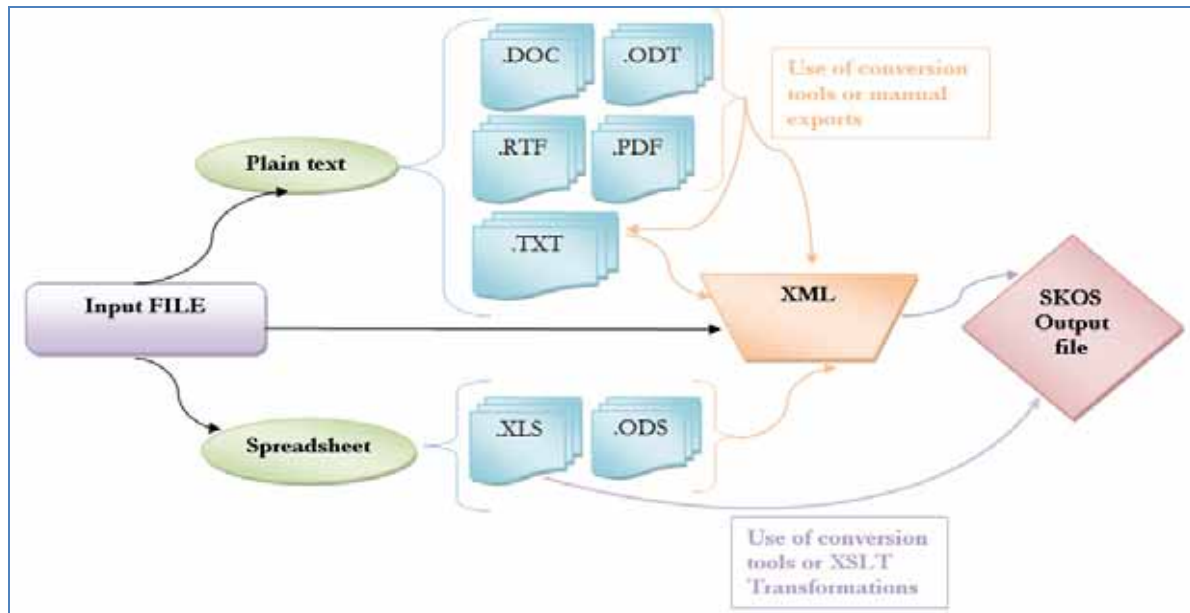


Figure 3: Transformation cases for SKOSification

The input file to be converted, e.g. “SKOSified”, may be a file containing text (e.g. a Word document or a PDF file), a spreadsheet file (e.g. Excel), or a structured document in XML format. Text files in proprietary formats can be converted with specific editing and conversion tools in order to get an XML transition format. For Word documents a command-line tool “Antiword” can convert it into TXT or XML format. For a PDF file, a manual export can be made in order to get a TXT version. From this text version an XML transition format can be obtained via some scripts or manual transformations. For Excel files, there is a possibility with the latest versions of Microsoft Excel or OpenOffice Calc to save the file as an XML document; some other tools for conversion into XML exist but are often shareware tools for example: ConvertXLS¹ which can convert Excel files in any output format. Other XML dedicated tools can also proceed to the conversion: for example Oxygen². A very recent shareware tool, Altova MapForce 2011³ allow to convert Excel data into XML and proceed to data mapping. Some tools have been recently developed to convert directly from Excel format to SKOS format.

Whatever the input format is, the use of XLST transformations for converting the transition XML format into the adapted SKOS version cannot be avoided. In order to perform these transformations, the user have to know well the structure of the terminology and be able to map the elements of its terminology to the SKOS datamodel. Then a person with technical

¹ <http://www.softinterface.com/Convert-XLS/Convert-XLS-T.htm>

² <http://www.oxygenxml.com>

³ <http://www.altova.com/download-excel-mapper.html?gclid=CJLv7InOhKQCFQYf3wodjjGiGw>

skills especially a good knowledge of XML and XSLT languages would implement the mapping set by the user to proceed to the SKOSification.

Till now we identified these situations, but others could be experimented with in the future, and other ways to get the final result may be found. To experiment with all the possibilities is important since there is no ideal method for all the situations. A case-by-case approach is necessary.

4.1.2. Mapping use cases

You work for a European museum, and you are in charge of the mapping of the terminology used in your institution with a reference thesaurus compliant with Europeana requirements. Your in-house terminology is already registered in a dedicated repository, and it is also SKOSified.

Main case:

In order to map your terminology with the reference thesaurus, you have browsed all the lists of terms proposed by the thesaurus. Among them, you are interested in some terms for your first mapping.

You identify the terms of your in-house terminology that you would like to map with those of the reference thesaurus, and you relate them. You define the type of semantic relationship between the terms according to the mapping format which enables to set equivalencies as exact, close or narrower/broader match. If necessary you make explicit the purpose of this relation in order to disambiguate.

Alternative 1:

You are already registered and SKOSified in-house terminology has previously been mapped with the reference thesaurus. You intend to change or to create relations to refine the alignment. As in the main case you can browse the thesaurus lists of terms and the terms of your in-house terminology. Most of all you can consult and edit the current version of the mapping, that is, the details of the relations.

Alternative 2:

Your in-house terminology has already been registered and SKOSified, and you work on its mapping the reference thesaurus. After a search in the thesaurus lists of terms you do not find the suitable terms to map to your own. Therefore you would like to enrich the reference thesaurus by proposing terms. In other words, you are going to propose an update of the reference thesaurus. The reason for the proposed update must be explained. In this case the mapping of your terminology is one step of an iterative loop for the refinement of the reference thesaurus. If your proposal is agreed to, the new version of the reference thesaurus will have the terms you proposed in it. You can now relate your terms with the reference thesaurus in an effective manner, especially if these new terms come from your in-house terminology.

4.2. Benchmark Results

In parallel to the setting up of a first version of the ATHENA Thesaurus, we made a benchmark of all the existing initiatives that have at least some of the functions of the 6 steps of the process of terminology management. This benchmark can be summarized in the following figure:

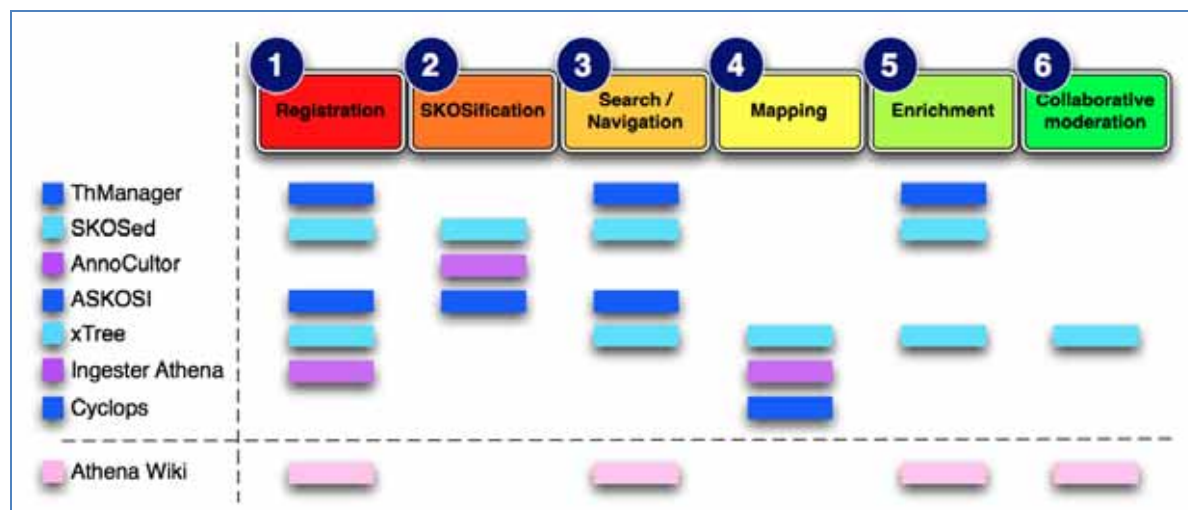


Figure 4 : Benchmark

ThManager¹ is an editing tool for SKOS thesauri which allow to registrate in an internal database several thesauri.

SKOSed² is a plug-in for the Protege software which is an ontology management tool. Annocultor³ is a set of command-line tools which allow to SKOSify a large number of terminology resources.

xTree⁴ is a web service for editing and enriching a SKOS terminology.

The Athena Ingestor⁵ presented above doesn't manage terminology but present interesting and useful features for terminology management.

Cyclops⁶ is a graphical mapping tool that could be adapted in order to map terminologies.

As we suspected, our benchmark confirmed there is no ideal tool for terminology management, i.e.:

- Able to cover the whole process (the six steps);
- Manage a user-friendly GUI and a powerful engine for display, search, and edition;
- Manage the gap of skills in-between those of the Museum people and those required by information engineering.

Therefore we can only say:

¹ <http://thmanager.sourceforge.net>

² <http://code.google.com/p/skoseditor/>

³ <http://annocultor.eu>

⁴ <http://www.digicult-sh.de/>

⁵ <http://athena.image.ntua.gr/athena>

⁶ <http://www.utc.fr/caspar/wiki/pmwiki.php?n=Main.Proto>

- We can recommend tools and methods separately for each step
- We can imagine for the future a complete environment integrating the whole process

For the SKOSification process, which is the main concern of this deliverable, we note that there is no generic tool that can convert into SKOS of any input file. This process has to be managed in a very technical way and knowledge and technical skills of a computer engineer may be required.

N.B.: All the details concerning main tools for terminology mapping and SKOSification are in the Annexes.

4.3.Means

In order to support our experiment we used the following means:

4.3.1. Working group activity

All the activity of the work package is supported by a working group (WG4). Mixing ATHENA partners and external experts, the working group takes advantage of the knowledge of the project issues in terms of usability in European museums, and the skills of experts in the domain of information engineering.

After the finalisation of the deliverable D4.1 the working group has met during three events¹:

- Budapest technical meeting – 13th November 2009
- Berlin technical meeting –25th February 2010
- Paris workshop – 25th June 2010

Budapest technical meeting²

A technical meeting took place in Petofi Museum (Budapest), the 13th of November 2009. This provided an opportunity to present the collaborative tool we set up for helping with work group activity (see below the specific part about the ATHENA Wiki).

The work group validated the work done on the process of terminology management for European museums and the list of related issues. Finally we benefited from the feedback of external experts about our approach. These experts are involved in different important projects or structures related to our topic: European Film Gateway³, Multimatch⁴, Europeana v1⁵, Europeana Connect⁶, Europeana Office⁷.

¹ All the information is available at:

http://www.athenaeurope.org/athenawiki/index.php/Documents#Meeting_and_working_documents

² http://www.athenaeurope.org/athenawiki/index.php/Documents#WG4_Second_meeting_.28Budapest.29

³ <http://www.europeanfilmgateway.eu/>

⁴ <http://www.multimatch.eu/>

⁵ <http://version1.europeana.eu/>

⁶ <http://www.europeanaconnect.eu/>

⁷ <http://www.europeana.eu/>

Berlin technical meeting¹

An informal technical meeting took place in Berlin the 25th February 2010, just before an ATHENA plenary meeting.

We used that slot for giving to the working group some specific feedback about the project review², and give information about the tasks to be achieved by WP4 in the following months.

During this meeting we also submitted for validation the format that has been set up for the ATHENA Thesaurus. This format is available via the ATHENA Wiki and is also detailed in the next section.

Finally a tool dedicated to the management of SKOS terminology, xTree, developed by the Digicult³ project was presented in the framework of the benchmark.

Paris technical workshop⁴

A whole day workshop took place in the French Ministry of Culture and Communication (MCC), Paris, the 25th June 2010.

We provided to the working group a set of presentations of different tools possibly compliant with the needs we identified about terminology management. Then we summarized the results of the benchmark we began work on several months before. Finally we benefited from the experience of external experts about the tools and the methodology. These experts are involved in different important projects or structures for our topic: EuroVoc⁵ SKOS model, ASKOSI⁶, the MACS⁷ project, SKOSification of the French thesaurus W for Archives⁸.

4.3.2. A collaborative environment

We have used a wiki for several months as a collaborative environment in which the members of the working group could (and still can) find back information and officially discuss ideas and results. We chose a wiki because it is quite easy to install on a server, and rather simple to manipulate as a contributor.

Technically speaking we decided to use Mediawikias wiki engine for two main reasons. First it is widely used (e.g. in Wikipedia) and several plug-ins could help us in our task. The plug-in *Semantic MediaWiki* allows for management of all the content of the Wiki in a Semantic Web oriented way. In addition, we also integrated Halo, which is a “Semantic plug-in” enabling the user to put categories and properties on pages as tags and annotations and providing a graphical interface for managing these annotations. Such a feature can be very useful in the collaborative activity as we imagined.

¹ <http://www.athenaeurope.org/athenawiki/index.php/Documents - WG4 Third meeting .28Berlin.29>

² The project review took place in Luxemburg the 2nd February 2010

³ <http://www.digicult-sh.de/>

⁴ <http://www.athenaeurope.org/athenawiki/index.php/Documents - WG4 Technical workshop .28Paris.29>

⁵ <http://eurovoc.europa.eu/>

⁶ <http://askosi.org/>

⁷ <http://www.athenaeurope.org/getFile.php?id=662>

⁸ <http://www.athenaeurope.org/getFile.php?id=661>

The wiki is available online at without any restriction:

<http://www.athenaeurope.org/athenawiki>

Contributors are invited to create a user account to be able to make any modification.

4.4. ATHENA Format

The ATHENA Format is the format in which the ATHENA Thesaurus is expressed. The WP4 working group agreed on the methodology and the format that were presented during the technical meeting in Berlin (February 2010) for the elaboration of the ATHENA Thesaurus. This format is proposed to the museums who want to map their own terminologies with the ATHENA Thesaurus. In this case, they have to use the ATHENA Format in order to form their descriptions before mapping. As a SKOS-compliant format, the ATHENA Format guarantees to the museums that their descriptions meet the Europeana requirement regarding SKOS.

Although SKOS is a basic structure for the formal representation of controlled vocabularies, it can be extended and customized very easily to have a more precise description of the terms and also include lexical elements related to these terms. The ATHENA Format is mainly based on the SKOS core data model, and it has been inspired by the museumvok format¹ in order to include some specific details.

4.4.1. Metadata

The metadata part of the ATHENA Format is intending to provide administrative information on the terminology that has been converted in SKOS.

Metadata	dc:title
	dc:creator
	dc:contributor
	dc:description
	dc:source
	dc:language
	status

These elements are borrowed from the Dublin Core data model (with the prefix “dc:”) and provide details about the terminology. Designing the metadata of the terminology in Dublin Core could eventually enable the OAI harvesting in the context of a repository or database of lexical or terminology resources.

¹ <http://museum.zib.de/museumsvokabular/index.php?main=tech-dok&ls=9&co=we>

The last element is not part of the Dublin Core data model but may be useful to check if the SKOS version of the terminology has a ‘valid’, ‘in validation’ or ‘draft’ status.

This set of elements is defined in order to get the same information for all the terminology resources that will be transformed into SKOS within the ATHENA project.

4.4.2. Concept

Concept	skos:Concept
	skos:ConceptScheme
	skos:inScheme
	skos:hasTopConcept
	<i>skos:topConceptOf</i>

As we already said, the concept is the central element of the SKOS data model. The data model makes a distinction between classes and properties. The first items of the table above *skos:Concept* and *skos:ConceptScheme* are classes whereas the next items (*skos:inScheme*, *skos:hasTopConcept*, *skos:topConceptOf*) are properties.

The property *skos:topConceptOf* is set in italics because it is the inverse property of *skos:hasTopConcept* then duplication of these two properties for linking two same concepts is not useful. Therefore this property is optional. When a property is the inverse of another one it supposes that only the subject or the object of an assertion need to have the mention of the property. A same concept cannot have these two properties at the same time with the same object.

For example:

A *skos:hasTopConcept* B

B *skos:topConceptOf* A

These two assertions express the same information then it is possible to use only one of them and avoid duplication of information.

4.4.3. Collection

concept Collection	skos:Collection
	skos:member
	<i>skos:OrderedCollection</i>

skos:memberList

The class for ordered collection and the corresponding property are set in italics to highlight that this is a possibility offered by the SKOS data model but it has to be used only if the order of the concepts within the collection is really relevant.

As we intend to bring together very different terminologies with very different scopes, the notion of collection may be useful to set these concepts as groups within the ATHENA Thesaurus. Indeed, some terminologies are only used for indexing, others are designed to improve information retrieval. Some terminologies are aiming at professionals whereas others are reachable by general public. The notion of collection can help to bring consistency among this diversity and give a facility to create thematic groups.

4.4.4. Description

concept Description	skos:prefLabel
	skos:altLabel
	skos:hiddenLabel
	<i>skos:notation</i>
	skos:changeNote
	skos:definition
	skos:editorialNote
	skos:example
	<i>skos:historyNote</i>
	<i>skos:note</i>
	skos:scopeNote

In this description block, we include the three different types of lexical labels. The preferred label is set in bold font because we define it as a mandatory property for the ATHENA Thesaurus. As we saw in the SKOS section, the SKOS data model does not force the use of labels for expressing a concept since a concept can be defined only through its semantic relations. But in the context of the ATHENA Thesaurus which is made from existing thesauri, the migration from descriptors to labels should be done carefully. Then we consider that the use of a preferred term is mandatory. We define the *skos:notation* property as optional since we gathered very few classifications during our inventory phase and therefore we privilege the use of labels instead of notations.

skos:note is the most generic type of note, then in order to force a more precise description of terms we set this property as optional in the ATHENA format.

skos:historyNote is mainly dedicated to keep track of diachronic evolution of terms. As the terminology resources gathered for the ATHENA thesaurus don't provide this information in

most of the cases, we set this property optional. Also, there might be a confusion between the *skos:historyNote* and the *skos:changeNote*; the *skos:changeNote* is mainly used to keep track of the evolution of description of a concept, e.g. a change in the labels used to express this concept or a change in its semantic relations.

Almost all the documentation properties have been included in the ATHENA Format since it is important to keep as much as possible of the information from the source terminology in order to keep track of the versions and changes of concepts.

As recommended by the SKOS data model, the language tags introduced in RDF by the @xml:lang attribute, are set as mandatory in the ATHENA Format in order to enable the multilingualism and highlight the linguistic richness of the resources that will compose the ATHENA Thesaurus. This attribute will be used for the labels and the documentation properties as well.

4.4.5. Relation

concept Relation	skos:broader
	<i>skos:broaderTransitive</i>
	skos:narrower
	<i>skos:narrowerTransitive</i>
	skos:related

These semantic relations constitute the core and the strength of the SKOS data model, and then it is logical to emphasize them in the ATHENA Format. Although the transitive properties *skos:broaderTransitive* and *skos:narrowerTransitive* are set in italics, since they may be useful to make transitive assertions, the use of these properties is optional.

4.4.6. Mapping

concept Mapping	skos:broadMatch
	skos:closeMatch
	skos:exactMatch
	skos:narrowMatch
	skos:relatedMatch

As for the semantic relations, the mapping properties constitute the essence of the SKOS data model. Then these properties will be used in the ATHENA Format to make alignment links between concepts from different concept schemes.

This section presented the format that will be used for all the terminology resources gathered during the inventory phase initiated for the first WP4 deliverable in order to constitute the

ATHENA Thesaurus. We wanted here to emphasize that the main features of the SKOS data model are reused in this format but although some of these features are made mandatory or optional in the framework of the ATHENA Thesaurus in order to get a homogeneous description of very heterogeneous terminologies.

4.5. Experiment results

4.5.1. Methodology

After surveying the terminology resources in use in European museums for the D4.1, an analysis of these resources has been made in that framework. All the in-house terminologies were gathered and have been organised according to a set of criteria:

- if the terminology is a thesaurus
- if the domain is specialized
- if the the terminology is multilingual
- if it is SKOSified

All the resources that are free of rights have been listed in the dedicated Wiki page.

Considering the huge diversity of terminology types, languages and subjects, we selected three resources in order to do a first mapping between them and then build a first core of the ATHENA Thesaurus as a first version.

The three selected resources are very different from each other but there is still an intersection from the subject and language point of view:

- Michael Terminology lists (Europe) available in 12 languages
- PICO thesaurus (Italy) available in English and Italian
- RMCA thesaurus (Belgium) available in French, English and Flemish

Almost all these terminology resources deal with cultural heritage or culture in its largest scope. Therefore the granularity and precision of each resource is very different from one to the other. Considering this, in order to test our guidelines and our final recommendations, we selected the three resources for their specific properties.

We checked if each of these resources was “ready to map”, that is:

- If it was SKOSified
- If it was proposed with persistent identifiers (URLs/URIs)
- If it was well-licensed for what we intended to do with
- If for that resource we could work with active contributors

At the beginning of the experiment the state was:

	SKOS	Persistent identifiers	License	Contributors
<i>PICO thesaurus</i>	OK	OK	OK	Giuliana De Francesco (MiBAC) and Karim Ben Hamida (MiBAC).
<i>Michael lists</i>	To do	To do	OK	Marie-VéroniqueLeroi (MCC)
<i>RMCA thesaurus</i>	To do	To do	OK	Roxanne Wyns (RMAH-KMKG)

At this point our experiment consists of achieving the end of the following 3-steps process:

1. *SKOSification*: We contribute to the SKOSification of Michael lists and RMCA thesaurus.
2. *Mapping*: MICHAEL – RMCA + RMCA – PICO (specific domain of architecture as an intersection of the three resources)
3. *Validation*: Three levels of validation:
 - a. Technical validation
 - b. Validation from the contributors of the SKOSified version of the resource
 - c. Validation from the contributors of the mapping links of this resource for the elaboration of the ATHENA Thesaurus.

This work respects the specific format based on the SKOS core data model and inspired by the museumvok format (see above the section ATHENA Format), and the result relates with LIDO. The domain “Architecture” was selected as an intersection of the three selected resources and is especially interesting because the precision level of description of the terms and concepts is very different from one resource to another one. Michael subject lists are very generic whereas the PICO thesaurus is very specific. Therefore this domain is very interesting from the mapping process point of view for the purpose of he experiment.

4.5.2. Core resources

Here are, briefly presented, the three main resources we used for building the first version of the ATHENA Thesaurus.

MICHAEL Terminology lists

Title	Michael Terminology lists - Subjects
Kind of resource	Thesaurus
Country	International
Language(s)	English, Czech, German, Estonian, Greek, French, Italian, Latvian, Dutch, Finnish, Bulgarian, Hungarian, Swedish, Slovenian, Spanish, Polish

Description	Terminology lists with thesaurus features for describing collections of the cultural heritage field. <u>Domain:</u> general Based on the Unesco thesauri for the subject headings. Mainly use for indexing and web browsing of the collections on the portal. Available in XML format.
Dimension	501-1000
URL	http://www.michael-culture.org/software/lists.zip

The table above provides a short introduction to the Michael Terminology lists. Several lists are available for subjects indexing but temporal and spatial indexing as well. For the purpose of the experiment we only focused on the subjects thesaurus.



Figure 5: Transformation of the Michael Subjects list

While carrying out the SKOSification of this terminology we noticed the following features and observations:

- The thesaurus has a hierarchy of three levels; each concept is introduced with the <item> markup and each term is introduced with the <label> markup; the level of hierarchy is indicated with the attribute “depth” which value is “1”, “2” or “3” according to the level of hierarchy.
- There are 191 concepts in the terminology
- Issue: Some concepts have the same label as their broader concept

PICO Thesaurus

Title	PICO thesaurus
Kind of resource	Thesaurus
Country	Italy
Language(s)	English, Italian

Description	<p>Thesaurus developed by the Italian Ministry of Culture - Ministero per i beni e le attività culturali (MiBAC)</p> <p>Mainly use for indexing and web browsing of the collections on the CulturaItalia portal.</p> <p><u>Domain:</u> Italian culture, with special focus on tangible and intangible heritage; people and organisations involved in cultural processes and administration; cultural and educational disciplines; chronological periods</p> <p>Available in SKOS format.</p>
Dimension	501-1000
URL	http://www.culturaitalia.it/pico/thesaurus/4.1/thesaurus_4.1.0.skos.xml

The PICO Thesaurus has the following features:

- The thesaurus is already in SKOS format. All the concepts are identified with Persistent URIs (PURL Persistent Identifier system, a short description of these systems is given in the Guidelines section)
- The thesaurus is organised in 4 concept schemes corresponding to the 4 thematic questions “who”, “what”, “where” and “when”.
- All the concepts are well documented with scope notes on the use of the labels and concepts according to the language.

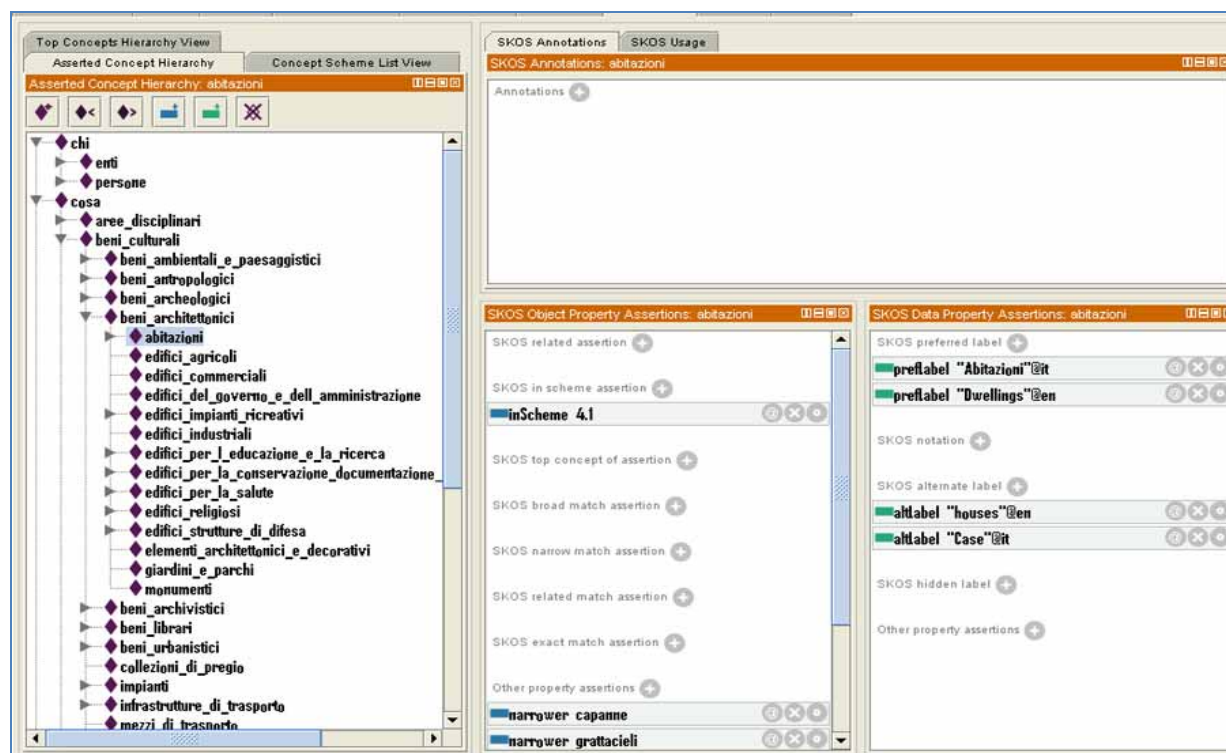


Figure 6: Preview of the Pico Thesaurus in the SKOSed (Plug-in for Protege) tool

RMCA Thesaurus

Title	RMCA Keywords
Kind of resource	Thesaurus
Country	Belgium

Language(s)	English, Dutch, French
Description	Thesaurus managed by the Royal Museum for Central Africa (RMCA) <u>Domain:</u> History and Ethnography Available in Excel format.
Dimension	11-100
URL	N/A

The RMCA Keywords thesaurus is an in-house terminology that is not available online. Here are the main features observed during the conversion:

- The thesaurus has two levels of hierarchy
- The distinction between the different level of hierarchy is done using cell colours
- The thesaurus is perfectly multilingual (each concept is expressed with a label in the three languages), however some terms are not available in Dutch and the English version is used instead. A validation process is ongoing in order to define if it is a coined term, e.g. a term that has been used because the corresponding language was not available or if the English version is the common use for this concept in that domain.

FACET THEMATIQUES (F)	FACET THEMATIQUES (NL)	FACET THEMATIQUES (E)
architecture/construction	architectuur/bouwwerk	architecture/construction
construction/chantier	bouw/bouwwerf	construction/work area
habitations/maisons	woningen/huizen	dwelling/houses
huttes/cases	hutten	huts
villes	stad	town/city
bâtiments	gebouwen	buildings
monuments	monumenten	monuments
meubilier urbain	straatmeubilair	street furniture
événements	gebeurtenissen	events

RMCA keywords
Excel

```

<rdf:Description rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#architecture">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:inScheme rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords"/>
  <skos:prefLabel xml:lang="fr">architecture</skos:prefLabel>
  <skos:altLabel xml:lang="fr">construction</skos:altLabel>
  <skos:prefLabel xml:lang="nl">architectuur</skos:prefLabel>
  <skos:altLabel xml:lang="nl">bouwwerk</skos:altLabel>
  <skos:prefLabel xml:lang="en">architecture</skos:prefLabel>
  <skos:altLabel xml:lang="en">construction</skos:altLabel>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#construction"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#dwelling"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#huts"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#town"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#buildings"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#monuments"/>
  <skos:narrower rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA_Keywords#street_furniture"/>
</rdf:Description>
  
```

RMCA keywords
SKOS

Figure 7: Transformation of the RMCA Keywords Thesaurus

The figure above shows the RMCA thesaurus in its source format and its converted form.

4.5.1. ATHENA Thesaurus V1

It is important to keep in mind that the ATHENA Thesaurus in its current version is at a draft status since it was mainly created for the purpose of testing in the experimental framework. So at this current stage, this thesaurus does not intend to be a standard.

Considering the properties of each of the selected resources, the elaboration of this first version of the ATHENA Thesaurus was done in two mapping steps.

The mapping was performed from the most general resource to the most specific. Then a first mapping between the Michael Subjects thesaurus and the RMCA keywords thesaurus was done before mapping this version with the PICO thesaurus which is the most specific.

The approach adopted to build this version of the thesaurus consisted in merging the non-published resources and make mapping links to the published one. We have considered each of the source terminology as a concept scheme.

In order to provide a thematic organisation of the concepts, and as designed in the PICO thesaurus, we set four thematic collections, namely “who”, “what”, “where” and “when”.

Here follows a screenshot of the ATHENA Thesaurus:

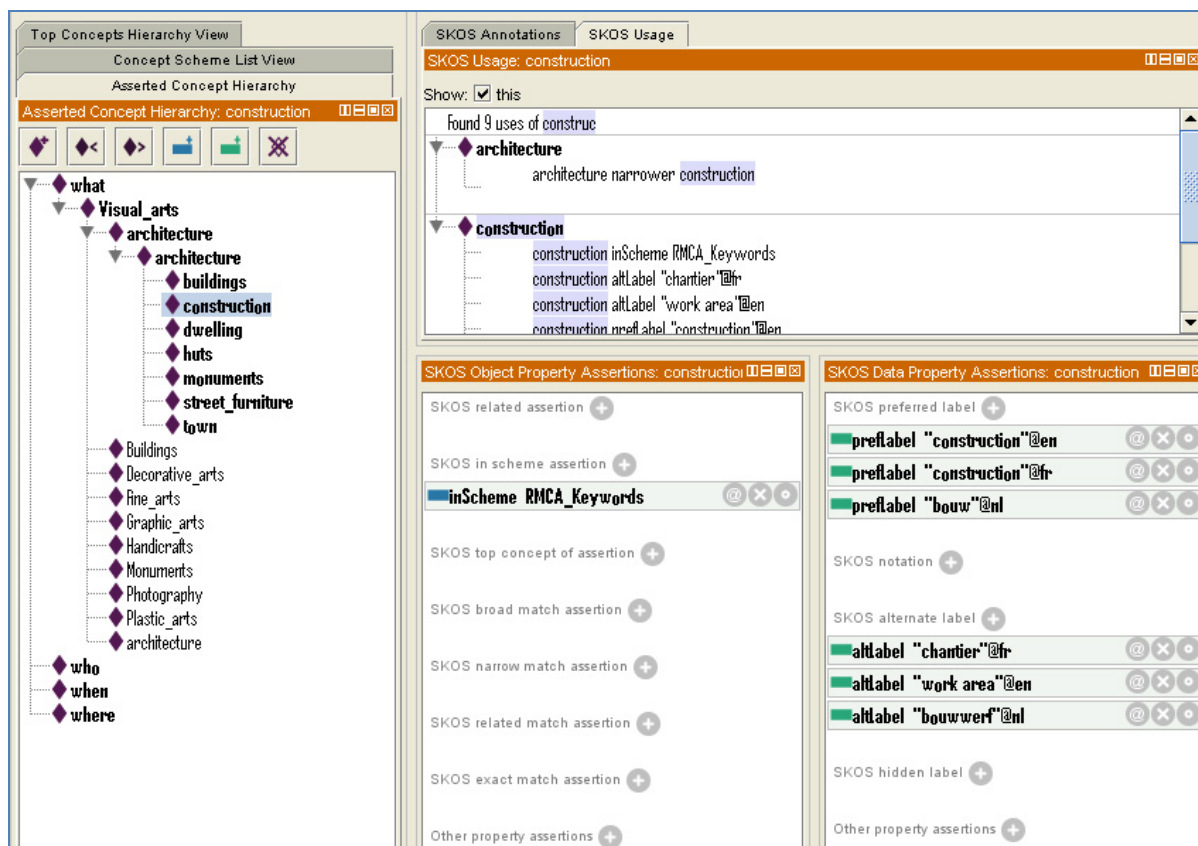


Figure 8: Preview of the ATHENA Thesaurus in SKOSed

The mapping of these three resources was done manually for the purpose of testing. As our benchmark on terminology and dedicated tools is still ongoing, tools for automatic or semi-automatic mapping will be studied.

In order to use the SKOSed tool, which is a plug-in to Protege (tool dedicated to the management of OWL ontologies), the URIs have been set in an explicit form. As we recommend in our guidelines. A next step for the elaboration of this thesaurus will be to choose a Persistent Identifier System (described in the Guidelines section) and define a sustainable way to identify the concepts.

5. Guidelines

The next deliverable (D4.3) will provide guidelines and general recommendations to the institutions concerning all the steps of the main process described in the section above (registration, search/navigation, enrichment, collaborative moderation). In this section, based on the focus of this deliverable, guidelines are provided with regards to SKOSification and Mapping steps. They are mainly from existing recommendations for managing thesauri, the major requirements of the SKOS model and our experience (observations and tests) within the experiment described above.

5.1. Benefits in using SKOS

RDFS and OWL are the languages that have been formally defined for knowledge representation. SKOS is one language among this formal languages' family. The major difference is that SKOS has been designed in to model every type of controlled vocabulary. It can be used to represent a thesaurus as well as a classification or a subject headings list. Then it is a good compromise for the institutions who are using these types of resources, and who are willing to be compliant with the Semantic Web technologies without developing sophisticated ontologies.

The SKOS data model is consistent with the formal ontology language OWL. Therefore the migration from a SKOS version of a terminology towards a formal ontology in OWL can be handled without major difficulties.

Since the SKOS model is very simple, but still complete enough, the implementation of a SKOS version has a low cost for migration. As we made the distinction in the introduction, SKOS is not a formal knowledge representation. But for an institution managing simple list of terms, or classifications and thesauri in the best case, it would be extremely costly and time consuming to develop a formal ontology perfectly compliant with Semantic Web technologies (using OWL for example). Therefore SKOS provides a structure based on classes and properties which give a powerful data model for migrating and porting these terminologies towards Semantic Web technologies.

Institutions must keep in mind that the adoption of the SKOS model is not a total replacement of the data model in use in the institution but a format for publishing and reusing their terminology and for ensuring the portability of this terminology for a semantic interoperability. Indeed usually knowledge organization systems (KOS), e.g. controlled vocabularies and thesauri, are used for indexing, and then porting these KOS into SKOS would enable the use of these indexing KOS for retrieval as well.

However SKOS may not be the appropriate language for every type of controlled vocabulary. For instance, authority files which usually provide a list of persons cannot be migrated to a SKOS version properly since the scope of this type of terminology is real persons and not concepts. Another point is that the SKOS semantic relations properties cannot really apply to authority files since a person cannot be related to another one with hierarchical (narrower/broader) or associative (related) links.

5.2. Guidelines for SKOSification

By SKOSification, we mean the process of conversion or transformation of a terminology into SKOS. We list below some guidelines for proceeding to this conversion from a technical and organisation point of view.

From the technical point of view, many of the guidelines provided here are inherent to the SKOS model but a special attention must be paid to these points in order to enable the general consistency within the ATHENA Thesaurus.

5.2.1. Evaluate the main features of the terminology to be migrated

Before starting any procedure for converting a terminology into SKOS, the institution must have defined the purpose of its terminology (e.g. indexing and retrieval, only indexing, or only retrieval).

As a second step, and a consequence of the definition of the purpose, the institution must evaluate if SKOS is the appropriate format considering the content of its terminology. In the case of authority files for instance, SKOS may not be the most appropriate format. Here are some features that can help for this evaluation:

- Concepts: Is the terminology dealing with objects and abstract things that could be assimilated to concepts? Is the terminology dealing with persons?
→ if the terminology is dealing with persons and not objects or abstract things, a standard like FOAF (Friend Of A Friend)¹ would be more appropriate
- Semantic relations: Are the descriptors (then concepts) of the terminology can be linked together via semantic relations.
→ if the terminology only contain independent descriptors without any semantic relations, a SKOS modelization is not absolutely necessary, an RDF representation may be more convenient.
- Interoperability: Can the terminology be linked to another resource dealing with the same subject/domain or scope?
→ if the terminology can be linked to other resources, all the potential links should be considered before the transformation process in order to implement these links in a most efficient way.

5.2.2. Identify your concepts

The W3C define two main steps to proceed to the identification of concepts:

- Creating (or reusing) a Uniform Resource Identifier (URI) to uniquely identify the concept

¹ FOAF : <http://www.foaf-project.org>

- Asserting in RDF using the *rdf:type* property that the resource identified by this URI is of type *skos:Concept*

The publication “Cool URIs for the Semantic Web”¹ from the W3C gives some main guidelines for the definition of the URI’s. Different systems have been elaborated in order to define Persistent Identifiers which give sustainability for the identification of resources. Indeed a URI is more than a simple hyperlink, persistent identifiers are supposed to continue to provide access to the resource, even when it moves to other servers or even to other organisations.

Several standards have been developed in order to normalise the definition of URIs. We give below a short description of these main standards². The ATHENA WP3 will produce a more detailed document on that subject and the final recommendations expected from the WP4 will take into account the outcomes of that deliverable.

- PURL
- URN
- NBN
- ARK
- Open URL
- DOI

PURL: A PURL (Persistent Uniform Resource Locators) consists of a URL; instead of pointing directly to the location of a digital object, the PURL points to a resolver, which looks up the appropriate URL for that resource and returns it to the client as an HTTP redirect, which then proceeds as normal to retrieve the resource. PURLs are compatible with other document identification standards such as the URN.

URN: The URN (Uniform Resource Name) is designed to describe an *identity* rather than a *location*; for example, a URN may contain an ISBN (International Standard Book Number, used as a unique, commercial book identifier).

NBN: National Bibliography Numbers (NBNs) is a URN namespace used solely by national libraries, in order to identify deposited publications which lack an identifier, or to reference descriptive metadata (cataloguing) that describe the resources. These can be used either for objects with a digital representation, or for objects that are solely physical, in which case available bibliographic data is provided instead.

ARK: The Archival Resource Key (ARK) is a URL scheme developed at the US National Library of Medicine and maintained by the California Digital Library. ARKs are designed to identify objects of any type – both digital and physical objects. The ARK scheme encourages semantically opaque identifiers for core objects. Unlike an ordinary URL, an ARK is used to retrieve three things: the object itself, its metadata, and a commitment statement from its current provider.

¹ <http://www.w3.org/TR/cooluris/>

² <http://www.ariadne.ac.uk/issue56/tonkin/>

Open URL: An OpenURL contains resource metadata encoded within a URL and is designed to support mediated linking between information resources and library services. This standard is not *primarily* designed as a persistent identifier/resolver but is described as a metadata transport protocol.

DOI: The Digital Object Identifier (DOI) is an indirect identifier for electronic documents based on Handle resolvers. According to the International DOI Foundation (IDF), formed in October 1997 to be responsible for governance of the DOI System, it is a ‘mechanism for permanent identification of digital content’.

We can see from these short introductions that some of these standards are more adapted to specific field (for instance, URN and NBN are more adapted for the libraries), however standards such as PURL or DOI could be used for definition of URIs.

Use of a Persistent Identifying System for the definition of the URIs

As we described them above, we recommend the use of standards for the identification of the concepts. Indeed, as the identification of concepts is achieved with the definition of HTTP URIs, these URI must be declared to persistent identification systems such as PURL which is normalised. This will also be of a great benefit since it is location-independent, e.g. if the terminology is moved from one location (housing server) to another, the URIs identifying the concepts of this terminology will not have to be modified.

Use of non-explicit URIs

It is highly recommended to use non-explicit URIs in order to avoid the reuse of a same URI for identifying two different concepts. Indeed as natural languages are by definition ambiguous and polysemous, it is possible that two different concepts might have two similar labels. The use of explicit URIs supposes that the choice of one specific natural language has been made during the definition or the migration of the terminology which cannot be convenient in a multilingual context.

5.2.3. Define with precision the labels expressing concepts

Preferred labels must be unique within a concept scheme

As it is required by the SKOS data model, no two concepts from a same concept scheme should have the same preferred label in a given language. However as natural languages are highly polysemous and full of homographs, the SKOS data model does not forbid that one concept can have two same preferred labels in two different languages.

Each concept must be expressed with one preferred label per language (mandatory)

As we saw above, the SKOS data model does not forbid the absence of preferred label, but labels are meant to help the understanding and refining the meaning of a concept. This is especially true in a multilingual context and it is helpful for purposes of administration and maintenance. Therefore we recommend using one preferred label per language.

It is important to note that this also means that it is not possible to have several preferred labels in the same language.

Avoid the concatenation of several words for a same label

In order to get the most accurate description, we recommend avoiding several values as a preferred term. For example, double concepts such as “dwelling/houses” must be considered as two different concepts that are linked by a semantic relation. The use of scope notes can help to reinforce the closeness of these two concepts.

The link between the two terms must be defined in order to provide the best description. We can state that “dwelling” and “houses” are synonyms; then the double concepts can be modelled as follows:

Dwelling: preferred label and houses: alternative label

Another possibility in the case of double concepts is to model the two concepts as related concepts.

Privilege the use of the lemma for the preferred label and possibly the other labels

The preferred label should consist in a single word term or a compound words term in natural language. This means that no artificial word or code must be used to label a concept. Such code must be defined using the *skos:notation* property.

The lemma of a word represents its canonical form. We strongly recommend this form of terms to be used as preferred label. For instance, in English or in French, the usual form of a lemma in the case of nouns is the singular for the number and the masculine for the gender. For verbal forms, infinitive forms will be privileged. Thus the forms of terms should be based on the conventions in the languages involved.

If the concept is only expressed with labels in specific forms that do not correspond to the lemma, this must be documented via the documentation properties (*skos:note*, *skos:changeNote*, *skos:editorialNote* or *skos:historyNote*)

In the case of compound terms, if possible, the addition of adjectives or verbs to a noun phrase should be limited.

In the same spirit, the use of articles and prepositions should be avoided in order not to extend the length of the label. From the computing systems point of view, these guidelines can help the efficiency of a retrieval system.

Privilege the typography in use by convention in the languages involved

The labels should respect the typographical rules that are usually in use in the languages of the labels. For instance, in English all the words referring to a language or nationality starts with an upper-case character whereas in French, these words will be in lower case characters. Thus we recommend respecting the conventions that are in use for each language involved. Any exception to this guideline must be documented via documentation properties of the model.

5.2.4. Avoid the duplication of information

The SKOS data model consists of classes and properties as we saw above. Meanings are to be deduced by an efficient use of these properties. As some of the properties available in the SKOS model are proposed as pairs (inverse or symmetric), this supposes that the use of one property implies the opposite or the reverse. Therefore it is better to avoid duplication and not to repeat the same information in different ways. SKOS terminologies are processed by machines. So the less redundant information there is, the faster the results of a query can be retrieved.

The main properties to pay attention to in order to avoid duplication of information are:

Inverse properties:

The use of the *skos:broader* or *skos:narrower* property implies the inverse meaning. Asserting that A has a broader concept B implies that B has a narrower concept A.

This is true also for the *skos:broaderTransitive* and *skos:narrowerTransitive* property.

Symmetric properties:

The *skos:related* property is symmetric then if an assertion that A is related to B is made, there is no need to make the following assertion, B is related to A.

However there is a possibility to use an extension to the SKOS data model in order to remove the symmetry of a property if this creates confusion in the meaning of the concepts.

5.2.5. Provide precision to the semantic relations of your concepts

Non-immediate hierarchical relations

In some cases, semantic relations between concepts have to be described with precision in order to avoid a loss of meaning or information and also avoid designing information which will not make any sense. For example the *skos:broaderTransitive/skos:narrowerTransitive* pair of properties allows to describe with precision relations between concepts when two levels of hierarchy are impacted.

Then the use of these transitive properties is preferred in order to assert a non-immediate hierarchical relationship between two concepts.

Consistency of the semantic relations

In order to ensure consistency, mixing hierarchical relationships with associative ones should be avoided. For example, a concept A cannot be related to another concept B if this concept A is the narrower concept of a concept C. Therefore a special attention must be paid when designing the semantic relations between concepts.

5.2.6. Enable the multilingualism

Provide for each concept an equivalent label in the languages involved in your terminology

Special attention must be paid to the multilingual labels expressing the concepts. These multilingual labels must be defined in the correct way in the different languages of the terminology so that the equivalencies can be computed from the SKOS representation of concepts.

Use the same system of language tags for defining the language of label

There are several systems which are normalized and equivalent: for example the three tags “en”, “en-GB” or “en-Latn” are different language tag systems referring to one language which is the English from Great Britain in Latin alphabet. In the case of terminology where different languages of different alphabet are involved, the tag system “language-alphabet” (for example “en-Latn”) may be useful for providing more precision. We recommend using the same system of tags for every language attribute of the terminology.

In the case where a specific language tags system is not required, we recommend the use of the language systems defined in ISO 639-1¹ where the language tags are coded on two letters in lower case.

5.2.7. Ensure the documentation of concepts and the terminology

Provide documentation for each change that may occur to a concept and its labels

The SKOS data model provides number of documentation properties in order to refine the meaning of a concept or keep track of the changes on the label(s) of a concept and/or its meaning. For the purposes of administration and maintenance of the terminology, each change must be reported in the SKOSified terminology using change notes (*skos:changeNote*) or editorial notes (*skos:editorialNote*)

¹ http://www.loc.gov/standards/iso639-2/php/code_list.php; see the ISO 639-1 column.

Provide as much as possible documentation to concepts with scope notes

As mentioned above, documentation on concepts helps to refine the meaning of a concept. The use of scope notes (*skos:scopeNote*) can be very helpful in enabling a better understanding of the concepts with contextual information. Examples may also be provided via *skos:example* property.

Documentation of concepts is especially needed in the case of homographs/homonyms in the same language or different languages for the labels expressing the concept. Then scope notes and examples can provide the user with a semantic disambiguation.

5.3.Mapping

Mapping is an inherent part of the SKOSification of a terminology. The following guidelines emphasize some aspects of the mapping process that may be crucial for general consistency of the terminology and the meanings of concepts.

5.3.1. Pay attention to the identification of your concepts during the mapping process

Use only absolute URIs

This guideline follows on from the one referring to the identification of concepts in the SKOSification part above. The terminology is made available in a machine-readable format by the SKOSification process. In order to make easily computable the identification of concepts and linking between concepts, it is recommended to use absolute URIs rather than relative ones.

For example:

`<rdf:Description
rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/RMCA
_Keywords#architecture">` is an absolute HTTP URI

`<rdf:Description rdf:about="RMCA_Keywords#architecture">` is a relative HTTP URI.

Respect the URIs of the original sources

As URIs are defined in order to identify the concepts uniquely, during the mapping process from a concept scheme to another, the URI defined within each concept scheme must be respected in order to enable the interoperability between the different resources involved.

5.3.2. Avoid the duplication of information

We saw that the structural properties for defining the semantic relations between concepts are either inverse or symmetric. This is also true for the mapping properties.

Inverse properties

The mapping properties *skos:broadMatch* and *skos:narrowMatch* are each other's inverse therefore there is no need to repeat twice the same mapping link using both properties for the same subject and object.

Symmetric properties

The mapping property *skos:exactMatch* and *skos:closeMatch* are symmetric. So repeating the mapping link can be avoided.

The property *skos:exactMatch* is also a transitive property then there is no need to repeat the mapping link on several levels.

For instance:

A *skos:exactMatch* B

B *skos: exactMatch* C

The assertion A *skos:exactMatch* C can be inferred from the preceding statement.

5.3.3. Provide precision to the semantic relations of your concepts

Use the appropriate properties to make links between concepts

The SKOS data model provides semantic relations and mapping properties, and does not restrict the use of these properties. However we strongly recommend to model in a homogenous way the relations between concepts in order to ensure the semantic consistency of the terminology.

We recommend to:

- Use mapping properties to make a link between concepts from different concept schemes
- Use semantic relations properties to make a link between concepts within a same concept scheme

The SKOS data model does not forbid using semantic relations properties for make a link between concepts from different concept schemes but it is highly recommended to follow these guidelines.

5.3.4. Enable the multilingualism

Manage multilingualism of the terminology through mapping of concepts and terms

The mapping process can be useful in a monolingual context but is especially relevant in a multilingual context. Equivalences can be stated from the mapping links made between several terminologies in different languages.

Equivalencies in a multilingual context can be of three kinds: semantic, cultural or structural. The semantic aspect refers to the meaning of the concept; the cultural aspect refers to the use of a term in a given language or culture; and the structural aspect refers to the semantic relations between concepts. This last aspect deals with the mapping and allows defining complete equivalence (synonymy) or partial equivalence (quasi synonymy) or non-equivalence.

As it was the case for the first version of the ATHENA Thesaurus, equivalences between concepts in languages that were not initially involved in the source terminology can be deduced from correct mapping links without translating the concepts.

5.3.5. Ensure the documentation of concepts and the terminology

Make explicit with notes the purpose of a relation

For the purposes of maintenance and administration, it is important to explain the choices of modelling that have been made for making links between concepts. The use of scope notes can help making explicit these choices.

Documentation properties can also keep track of history of mapping links.

Validation is an important part of the SKOSification process and mapping also. Therefore a special attention must be paid to this final step of the SKOSification.

From a technical point of view, in order to check the consistency of your converted terminology to the SKOS model, we recommend using the online web service Pool Party¹. Pool Party offers a free online tool for validating SKOS files that may be already online or stored on your local repositories.

This tool checks the consistency of the SKOSified terminology according to the following points which refer to our guidelines:

- Valid URIs: the tool checks if there is not any unauthorised character in the URI. Although if an URI is used twice for identifying two different concepts, there won't be any alert or warning.
- Missing language tags: the tool checks if all the labels and notes have a language tag
- Missing labels: the tool checks that each concept has at least one preferred label.
- Loose concepts: all the concepts that are isolated and not linked to other concepts are pointed out as loose concepts

¹ <http://demo.semantic-web.at:8080/SkosServices/check>

- Disjoint OWL classes: the tool checks the eventual consistency with OWL elements that may be in the SKOSified terminology
- Consistent use of labels: the rules for the use of labels are checked by the tool in order to avoid the use of a same label as a preferred label and alternative or hidden label, and to avoid the use of two preferred labels in a same language, ...
- Consistent usage of mapping properties: the tool checks the consistency in the mapping relations.
- Consistent usage of semantic relations: the tool checks that there is no mix between hierarchical and associative semantic relationships.

An example of output from this tool is presented in the Annexes.

From the content point of view, only the administrators and users of the terminology can validate the final migration of the terminology into SKOS format at least for an initial transformation process since they will be the one able to confirm or modify the general design of the terminology and its semantic relations according to the indexing and retrieval efficiency. For further modifications and updates, a set of rules and policies have to be defined in order to enable the collaborative moderation for managing the terminology. These rules and policies have to be agreed on by the community of users.

6. Conclusions

6.1. What would be an ideal tool?

This conclusion about an ideal tool comes from the needs and issues raised through the activity done in the Work package 4 of the ATHENA project. The work done in the WP4 framework is more and more making explicit unsatisfied needs for European museums and cultural institutions more generally in terms of terminology management and harmonisation of the existing terminology resources. For the time being, provision of content from European museums to Europeana is facilitated through ATHENA though the technical diversity and the multilinguality of their terminology make difficult their compliancy with Europeana requirements. There is a gap between the actual situation of terminology management in museums, and the skills and means necessary to have for an effective ingestion into the portal. Thus we look for one possible answer to reduce such a gap.

The needs:

When an institution intends to ingest into Europeana its digital collection and object descriptions, an effort about the terminology in use for these descriptions has to be done. Indeed, there is a set of criteria to respect in order to be compliant (SKOS format, multilinguality). This task of terminology management internally requires an expertise and tools that are not available in the institutions most of time. For instance, in ATHENA WP4, D4.1 study (Identification of terminology resources in European museums) has confirmed that a lot of European museums use an in-house non-standard terminology to describe their collections and objects. The cost implied by a reference terminology or specific needs (language, domain, ...) are the main reasons for this choice. This means that these museums have a strong effort to make for expressing their descriptions with a reference terminology fitting with Europeana. Particular skills in knowledge management and/or information engineering are necessary to have internally. Tools have to be identified, possibly acquired, and tested together to make sure that they are interoperable. Such an effort is very costly and time consuming for the museums.

Context assumptions:

We assume that:

- The European Commission cannot take in charge all the institution costs of terminology management for the compliancy with Europeana (this means that institutions will have to support such a cost we should try to reduce the more we can)
- The Web 2.0 approach and practice can provide new opportunities for terminology management at an institutional level and can make easier cooperation with the European level.

One possible answer:

The answer we propose consists of the design and the implementation of an integrated software environment for terminology management enabling any institution to find its way to manage its terminology according to Europeana ingestion.

A study of the information process (see the Annexes) concerning terminology management has already led us to consider an integrated software environment able to support a 6-step-chain of tasks:

1. Registration of a terminology in a repository
2. SKOSification of a terminology
3. Search and navigation into a network of vocabularies
4. Mapping of the terminology with a thesaurus
5. Enrichment of a thesaurus
6. Collaborative moderation of updates/modifications of the thesaurus

In our understanding, we consider that an ideal tool for terminology management would have as features:

- To be a **webservice**: For collaborative work online through Internet (e.g.: Athena Ingester service)
- To have a **user-friendly GUI**: Adapted for a non-expert use in European museums (e.g. Cyclops for graphical mapping)
- To combine **open-source** components: Such a service must stay independent of proprietary codes and formats (e.g: xTree)
- To be logically structured with an **intuitive Workflow**: The user must find which actions to do according to his/her needs(e.g.: WP4 6 steps process)
- To be **flexible** enough to be adapted to new standards: What if SKOS is updated in a new version or evolving towards an ontology description?

6.2. Perspective

After having provided a survey on the terminologies used in European museums (D4.1) and guidelines for SKOSification and mapping (D4.2), we are now going to deliver by the end of the project the final recommendations about terminology management (D4.3). In a certain sense, that coming deliverable can be considered as a specification report for the implementation of the ideal tool we mentioned above.

We can also guess that such an effort to specify the ideal tool may complement other initiatives or projects. For example we can mention a new project **Linked Heritage** (under negotiation), which addresses the coordination of Standards and Technologies for the enrichment of Europeana. Actually a WP dedicated to terminology management and multilingualism is foreseen in the project. The WP aims to take advantage of ATHENA WG4 activity, and to implement a prototype of an integrated software platform for terminology management. The WP will also benefit from the work done on the ATHENA Thesaurus in order to continue its completion and extend it to the other domains as Linked Heritage is a cross-domain project.

Moreover, as our benchmark helped us to identify relevant tools and structures in regards with terminology management, possible solutions and developments are to be investigated in order to adapt the ATHENA Ingester for terminology mapping or integrating components of the xTree tool. In both cases, collaboration with partners of the Linked Heritage project will be reinforced in order to propose a sustainable solution.

7. Annexes

7.1. Acronyms

ARK: Archival Resource Key
DOI: Digital Object Identifier
HTML: Hypertext Markup Language
HTTP: Hypertext Transfer Protocol
LIDO: Lightweight Information Describing Objects
NBN: National Bibliography Numbers
OWL: Web Ontology Language
PURL: Persistent Uniform Resource Locators
RDF: Resource Description Framework
RDFS: RDF Schema
SKOS: Simple Knowledge Organisation System
SPARQL: SPARQL Protocol And RDF Query Languages
URI: Uniform Resource Identifier
URN: Uniform Resource Name
W3C: World Wide Web Consortium
XML: Extensible Markup Language

7.2. Process and issues

Here follows the content of a document we finalised 2nd November 2009 and that were presented during the technical meeting in Budapest the 13th of November 2009.

Athena WP4 Terminology process

Purpose

Here is the description of the logical processes concerning terminology within the project Athena. This document is a draft and its content shall both help write the WP4 deliverables and drive the experiment we are planning. So: For discussion.

Introduction

Here are a few remarks necessary to keep in mind by reading this document.

Museums

Among all the possible content providers we focus on the European museums:

- Which have descriptions of the objects composing their digital collections,
- Which have used a terminology to express these descriptions,
- Which have made available or intend to make available their repository for a harvesting by Athena, hence by Europeana.

Use cases

For such an institution, we have listed different scenarios that we can distribute into 2 major categories. The first main category takes into account 3 cases where the processes are fully achieved through Athena and Europeana without any update of the Athena Thesaurus. The second main category deals with 3 other cases where an evolution (modification/update) of the Athena thesaurus has to be achieved.

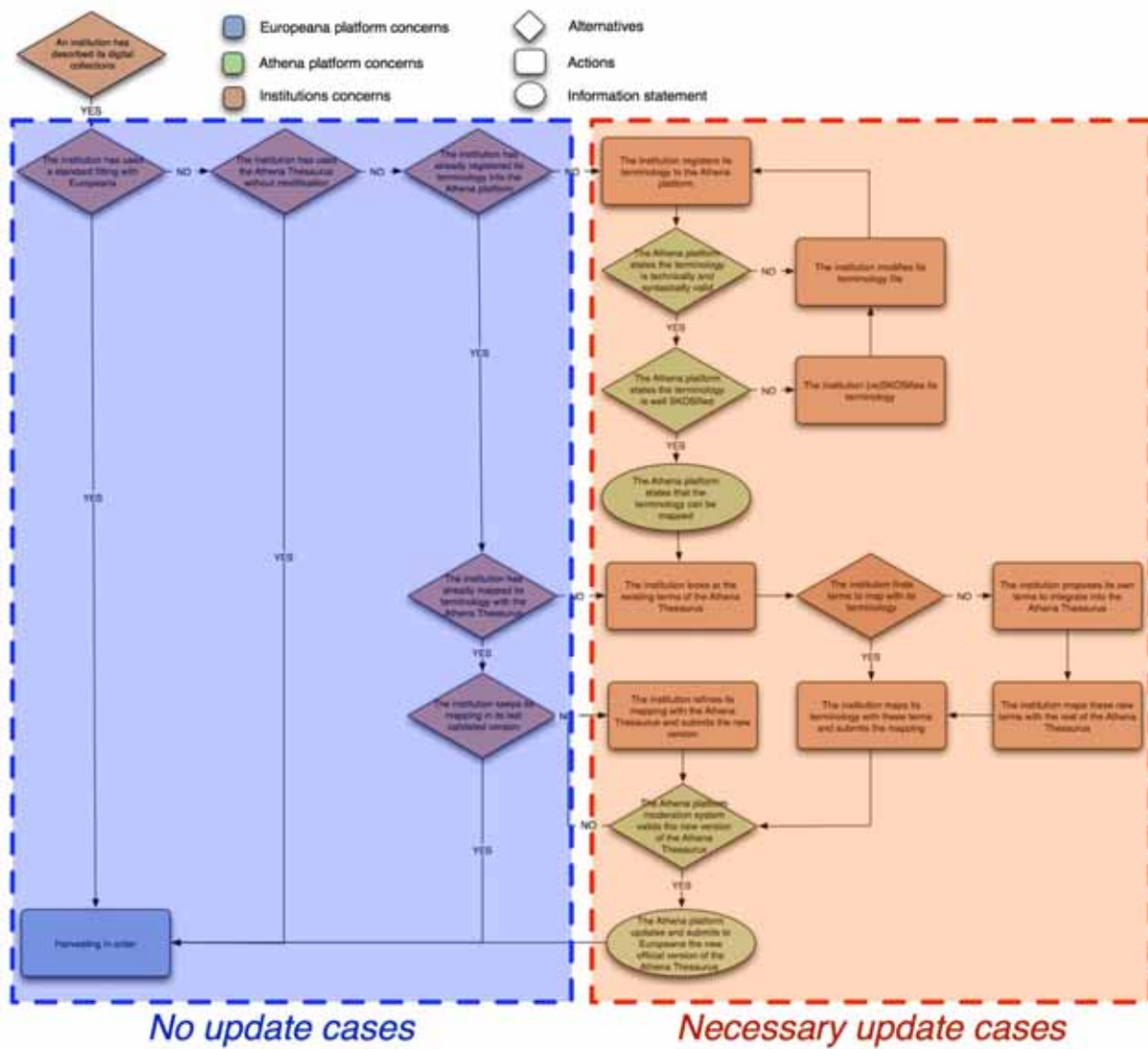
No update cases:

1. The institution has used a standard fitting with Europeana for describing its collections
2. The institution has used the Athena Thesaurus without modification for describing its collections
3. The institution has used another terminology which is already registered into the Athena Platform and mapped with the Athena Thesaurus, and does not aim to update anything

Necessary update cases:

4. The institution has used another terminology which is already registered into the Athena Platform and mapped with the Athena Thesaurus, and just aims to update its mapping
5. The institution has used another terminology which is already registered into the Athena Platform but which is not mapped with the Athena Thesaurus yet, and aims to do this mapping

6. The institution has used another terminology which is not already registered into the Athena Platform, and aims to register it and map it with the Athena Thesaurus



Athena Thesaurus

We call Athena Thesaurus the thesaurus produced and updated by all contributors during and after the project. As a thesaurus, the Athena Thesaurus is a network of controlled vocabularies, that is, an amount of terms organised by domains of description and structured thanks to bridges in-between.

This Athena Thesaurus is:

- **SKOSified:** The Athena Thesaurus is already SKOSified; it fits with Europeana requirements; so it can be directly used for description by institutions in case
- **Free of rights:** Any institution can use it as it likes without paying any fee; hence an institution which enrich the Athena Thesaurus by terms coming from its own terminology must check if it has rights to do so for free distribution and modification
- **Evolving:** We are considering to enable a collaborative workflow to produce and update the Athena Thesaurus; a specific interface with moderation process can be imagined
- **Available online:** We can imagine a Web service helping an institution to use the Athena Thesaurus online for description; of course this terminology will be downloadable for a use offline
- **Mappable:** We consider to enable the mapping through a Web service of terminologies with the Athena Thesaurus; to do so, there are a few requirements: 1/ the terminology must be syntactically and semantically valid; 2/ it must be well-SKOSified. While these requirements are not satisfied, the mapping would not be possible.

Principles

Updating

In a nutshell, hereafter we have represented the processes of these use cases according to the principle of updating. In this way the question we tried to answer step-by-step was: Is there someone aiming to create or modify something? Then we identified the possible “who” of each action among: the institution, the Athena platform, the Europeana platform. This must absolutely be discussed and validated, especially by the WP7 leader.

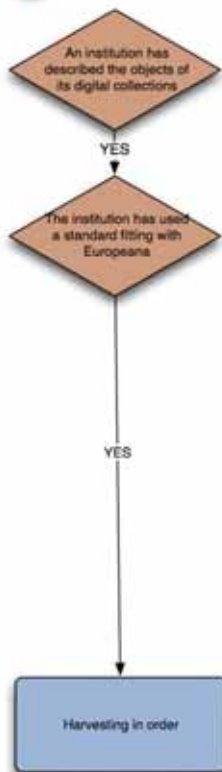
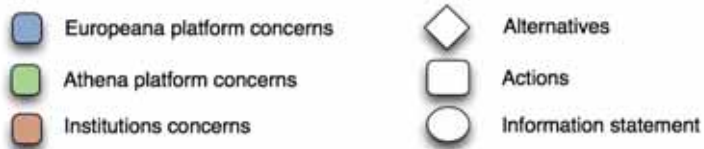
Any updating of the Athena Thesaurus is motivated by a change of the mapping between its elements. In certain case, this change is due to a new terminology that some institution has mapped with the Athena Thesaurus. And the mapping of a new terminology requires the validation of its form: the terminology must be well-SKOSified.

Processes

No update cases

Use case 1

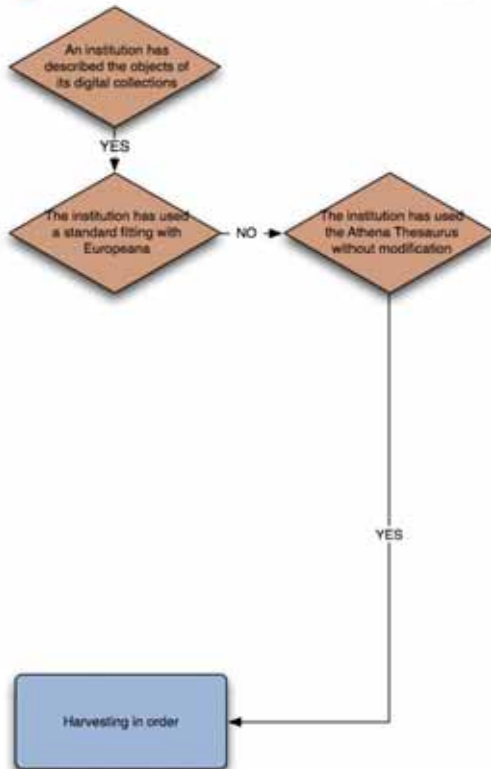
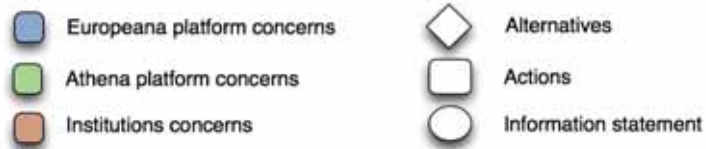
The institution has used a standard fitting with Europeana for describing its collections



This is the simplest case in the sense that it does not require any specific action from the Athena terminology scope. Indeed the institution has used a standard already compliant with Europeana requirements to describe the objects of its collections.

Use case 2

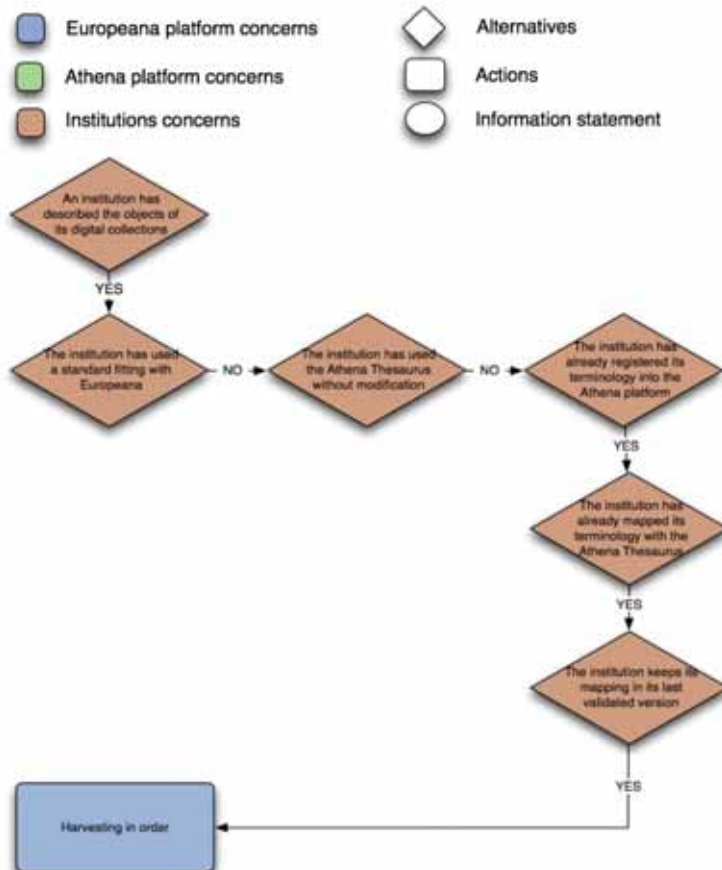
The institution has used the Athena Thesaurus without modification for describing its collections



This case is also very simple, like a particular case of the previous one. Here the institution has used the Athena Thesaurus to describe the objects of its collections. Now, as far as the Athena Thesaurus is used without modification in-house, it is compliant with Europeana requirements. This is one guarantee that Athena can provide.

Use case 3

The institution has used another terminology which is already registered into the Athena Platform and mapped with the Athena Thesaurus, and does not aim to update anything



Here is the last simple case we have identified. The institution has used another terminology to describe the objects of its collections rather than the compliant standards nor the Athena thesaurus.

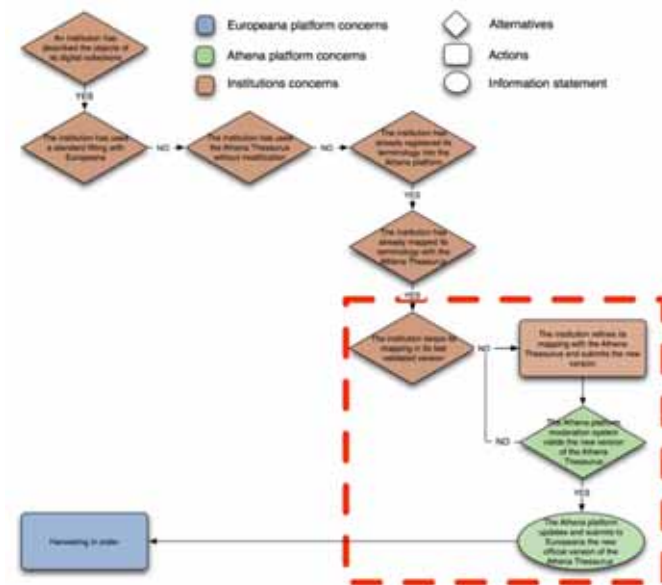
But it has already registered it into the Athena Platform and mapped it with the Athena Thesaurus. So all the descriptions expressed with this terminology are exploitable by Europeana for access and retrieval. This is due to the mapping with the Athena Thesaurus.

Since no update is foreseen, harvesting is in order as before.

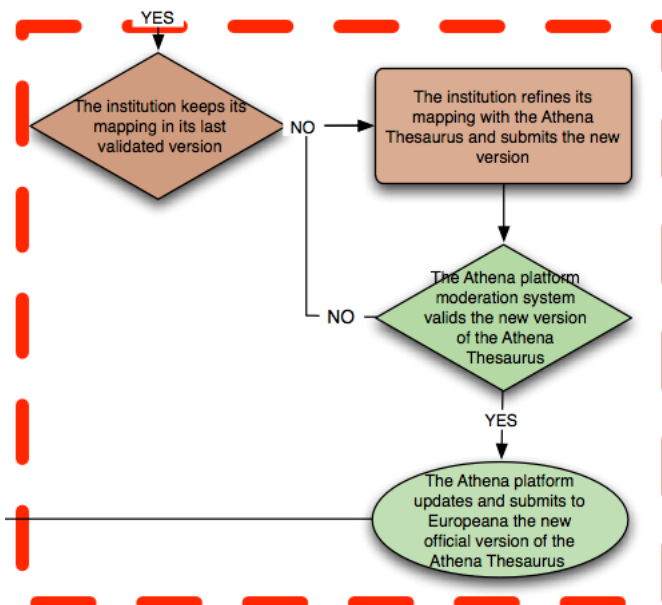
Necessary update cases

Use case 4

The institution has used another terminology which is already registered into the Athena Platform and mapped with the Athena Thesaurus, and just aims to update its mapping



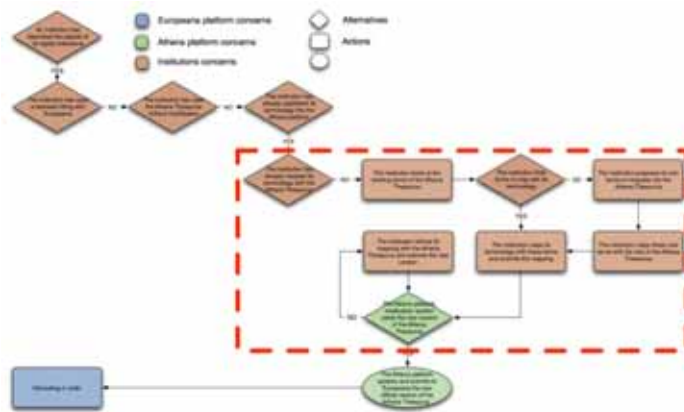
Here is the first case of updating of the Athena Thesaurus. The institution has used another terminology to describe the objects of its collections rather than the compliant standards or the Athena thesaurus. It has already registered it into the Athena Platform and mapped it with the Athena Thesaurus.



But the institution needs to refine the existing mapping. So the Athena Platform must check if the new mapping is correct. If yes, the Athena Platform submits the update of the Athena Thesaurus to Europeana, and harvesting keeps possible as before. If no, the institution must refine again and again its mapping until the platform validates the result. In case of cancellation, the last valid mapping version is still applied.

Use case 5

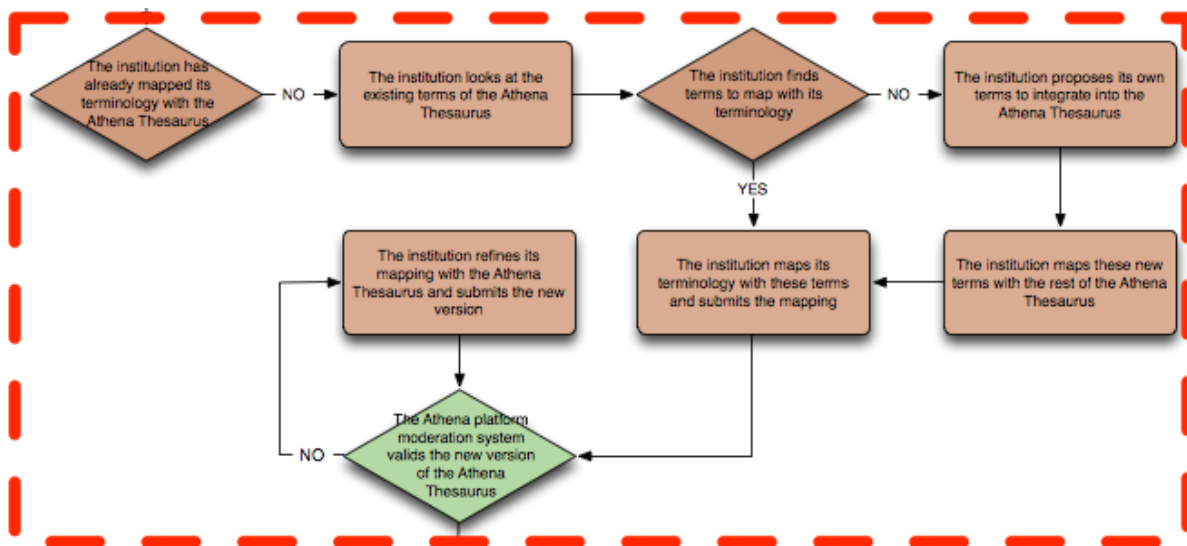
The institution has used another terminology which is already registered into the Athena Platform but which is not mapped with the Athena Thesaurus yet, and aims to do this mapping



Here is the case of updating of the Athena Thesaurus in which a first mapping has to be made with the institution terminology.

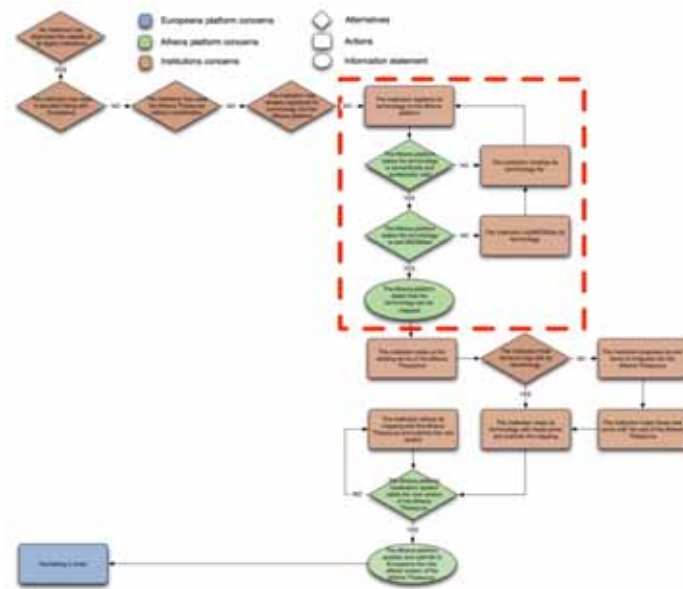
The institution has used another terminology to describe the objects of its collections rather than the compliant standards or the Athena thesaurus. It has already registered it into the Athena Platform, however it did not map it with the Athena Thesaurus yet.

To do so, the institution is invited to look at the existing terms of the Athena Thesaurus to find equivalent ones to its own terminology terms. Thanks to a domain organisation of the Athena Thesaurus and a graphical display of all its controlled vocabularies, the research would be more effective. If it finds relevant terms, the institution maps its terminology terms with those, then the Athena platform controls if the mapping is correct like it was a simple updating. If the institution does not find relevant terms, it can propose its own ones to enrich the Athena Thesaurus. Once again a validation process (with moderation) is necessary. The following of the process is like within the previous case.



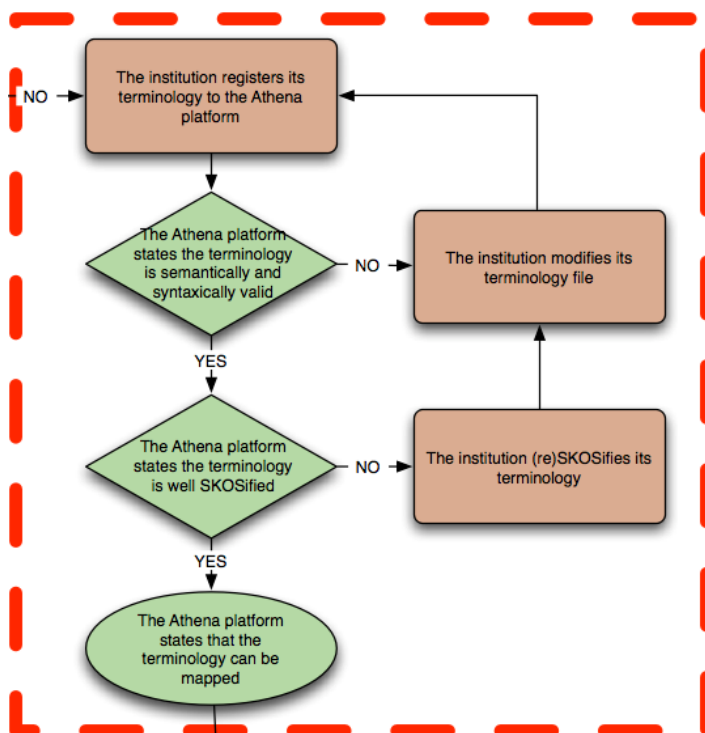
Use case 6

The institution has used another terminology which is not already registered into the Athena Platform, and aims to register it and map it with the Athena Thesaurus



Here is the final case of updating of the Athena Thesaurus in which a complete process of registration and mapping is necessary.

The institution has used another terminology to describe the objects of its collections rather than the compliant standards or the Athena thesaurus. It has not already registered it into the Athena Platform.



The registration is mandatory before any mapping with the Athena Thesaurus. This is a 2-step process.

First the Athena Platform checks if the terminology is semantically and syntactically valid (it means: if the file can be interpreted). Then it checks if the terminology is well-SKOSified.

If these two requirements are not satisfied, the Web service of the Athena platform does not allow the mapping.

If they are satisfied, the institution can map its terminology with the Athena Thesaurus as presented in the previous case.

Issues

Thanks to this process representation, we have listed for the time being 4 issues to discuss with Athena partners:

- **Workflow:** We are planning to deliver finally a workflow specification of collaborative production of the Athena Thesaurus; the moderation process appears as the tricky point of such a workflow;
- **Versioning:** We keep in mind that such a service must ensure at every moment that all previous versions of the Athena Thesaurus are still working for harvesting. Each update (modification/deletion/addition) should be detailed and archived.
- **Platform(s):** We wonder how to technically support SKOSification and mapping tasks: is it in the perimeter of the Athena platform? Shall we expect that Europeana will provide similar Web services (and hardware) we could duplicate? (→ To discuss especially with WP7 and EuropeanaConnect or v1.0)
- **IPR:** We consider to let the Athena Thesaurus free of rights for use and (controlled) modification; a Creative Commons “By:” license might be useful → To discuss especially with WP6
- **Sustainability:** What Athena and/or Europeana can ensure as a service after the project? What if we propose at the end of the project an Athena Thesaurus Web service for the online use of the Athena Thesaurus for description of collections?

7.3. Benchmark: Tools

All the tools and initiatives we have considered are listed in the table below:

NAME	TAGS	COMMENTS
NSDL Registry	methodology; graphical interface	
Chimaera	tool;	Chimaera was built on top of the Ontolingua Distributed Collaborative Ontology Environment.
SWOOP	tool; graphical interface	SWOOP is no longer under active development at mindswap. Continuing development can be found on SWOOP's Google Code homepage at http://code.google.com/p/swoop/
Visual Thesaurus	graphical interface	Proprietary and costing interface.
Jibiki platform		
VMF	methodology;	VMF means Vocabulary Mapping Framework - interesting here for its methodology, because it deals with data formats instead of terminologies.
eXo Platform	methodology; graphical interface	Interesting as a workflow management support
Visuwords	search; navigation; mapping; graphical interface; navigation map; search engine	available for reading online; can we use it as an editor?
Morphon XML Editor		XML editor sometimes used for Ontology expression (not very interesting for us)
SKOSed	graphical interface; skosification; search engine; classification	plug-in for Protégé
MultiTerm	graphical interface; mapping	Commercial Product for multilingual mapping of terms
PoolParty	skosification; format recognition; syntax checking	Conversion to SKOS works only from Zthes format
Os Meta Search	search engine	Meta search engine
XTree	graphical interface; registration; classification; account management	
iMap		
AnnoCultor	format recognition;	From databases and XML files to RDF; no graphical interface
ASKOSI	registration; format recognition; syntax checking; skosification; graphical interface	registration through DSpace.org; willing to guarantee the maintenance of a set of tools
XL2XML	skosification	Excel to SKOS/RDF conversion tool; no graphical interface
W3C SKOS Core Validator		based on PoolParty in its latest version
Cyclops	graphical interface; mapping; UGC management	Archive representation service for preservation
SKOS2OWL		Conversion tool from SKOS schemas to OWL

Figure 9: table of all the tools and initiatives considered in our benchmark. The picture is a screen shot of the one available on the wiki at: <http://www.athenaeurope.org/athenawiki/index.php/Benchmark#Tools>

Among all these tools we have particularly focus on the ones that have briefly presented below.

7.3.1. ThManager

Introduction

Name	ThManager
Main function	Edition and visualization
Administrative Information (Public/private, country, language)	Public (University of Zaragoza - GeoSpatiumLabS.L., Spain), English and Spanish
Type (tool, web service)	Standalone software
Command-line, GUI	GUI
OS	Multi-platform (Windows, Mac, Unix)
Skills and requirements	Installation of Java (JVM)

Remarks	Metadata on thesauri are managed in a separate file (DC)
---------	--

Benchmark outcome



- **Registration:** table of thesauri, then not a real navigation tree for the open thesaurus
- **Search / Navigation:** search engine inside one terminology
- **Enrichment:** Edition mode

7.3.2. SKOSed (Protégé)

Introduction

Name	SKOSed (Plug-in Protégé)
Main function	Ontology production
Administrative Information (Public/private, country, language)	Public (Stanford University), USA, English
Type (tool, web service)	Standalone software
Command-line, GUI	GUI
OS	Windows, Mac OS
Skills and requirements	Information engineering, grasp of Protégé
Remarks	No real SKOSification, starting from scratch; OWL compatibility, good for Semantic Web

Benchmark outcome



- **Registration:** Connection with online repository, local file system
- **SKOSification:** From scratch only, not user-friendly
- **Search / Navigation:** search engine inside one terminology, or navigation into a list of ontologies
- **Enrichment:** Edition mode

7.3.3. AnnoCultor

Introduction

Name	AnnoCultor
Main function	Conversion (XML to RDF)
Administrative Information (Public/private, country, language)	Public (Multimediane-Culture Project; Europeana Project), English
Type (tool, web service)	Set of tools
Command-line, GUI	Command-line
OS	Multi-Platform (Windows, Unix, Mac)
Skills and requirements	Installation of Java (JDK) and Apache Maven (software project management tool); Technical skills (XML, XSL and Java)
Remarks	Finished Project; conversion of collections and thesauri

Benchmark outcome



- **SKOSification:** Command-line

7.3.4. xTree

Introduction

Name	xTree
Main function	Production, edition and visualization
Administrative Information (Public/private, country, language)	Public (Digicult Project, University of Kiel); German
Type (tool, web service)	Web service
Command-line, GUI	GUI
OS	Multiplatform (Windows, Mac OS, Linux)
Skills and requirements	Knowledge <i>priori</i> of the terms to map

Remarks	Built on Open Source software, no flexible graphical mapping
---------	--


Benchmark outcome



- **Registration:** Connection with online repository, local file system
- **Search / Navigation:** tree, search engine inside one terminology, or navigation into a list of ontologies
- **Mapping:** on a same page with SKOS relations, but without any intuitive graphical interface
- **Enrichment:** Edition mode
- **Collaborative Moderation:** Forum

7.4.Pool Party SKOS validator

Here is the output of the SKOS validator of Pool Party



[Upload another file](#)

Results

<p>✔ Valid URIs: Passed! Checks if URIs are valid and do not contain any invalid characters like whitespaces.</p>
<p>✔ Missing Language Tags: Passed! Handles missing language tags of all sorts of SKOS labels and textual content.</p>
<p>✔ Missing Labels: Passed! Checks for missing labels - prefLabels for skos:Concepts and rdfs:labels for skos:ConceptSchemes.</p>
<p>✔ Loose Concepts: Passed! This checks handles loose concepts, i.e. concepts that are no topconcept in any scheme and have no rdfs:label.</p>
<p>✔ Disjoint OWL Classes: Passed! Checks if there are any instances of owl:Classes that are declared disjoint.</p>
<p>✔ Consistent Use of Labels: Passed! Checks if there are concepts with clashing SKOS labels, i.e.:</p> <ul style="list-style-type: none">• More than one prefLabel in the same language• A prefLabel that is also a hiddenLabel• A prefLabel that is also an allLabel• An allLabel that is also a hiddenLabel
<p>✔ Consistent Usage of Mapping Properties: Passed! Checks if concepts are connected by clashing SKOS mapping relations:</p> <ul style="list-style-type: none">• skos:exactMatch and skos:broadMatch• skos:exactMatch and skos:relatedMatch
<p>✔ Consistent Usage of Semantic Relations: Passed! Checks if concepts are connected by clashing semantic SKOS relations:</p> <ul style="list-style-type: none">• connected by skos:related and skos:broaderTransitive• connected by skos:related and skos:narrower

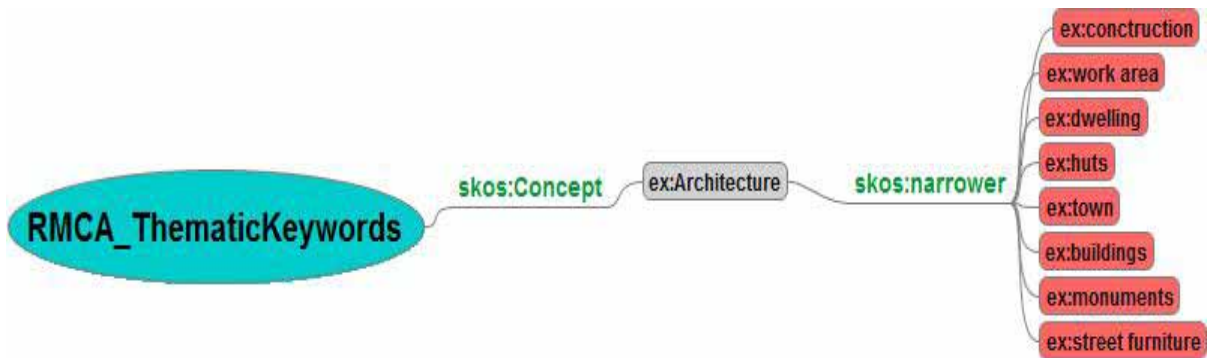
✔ Check passed
✘ Mandatory check failed (not consistent with SKOS specification)
○ Optional check failed

[Upload another file](#)

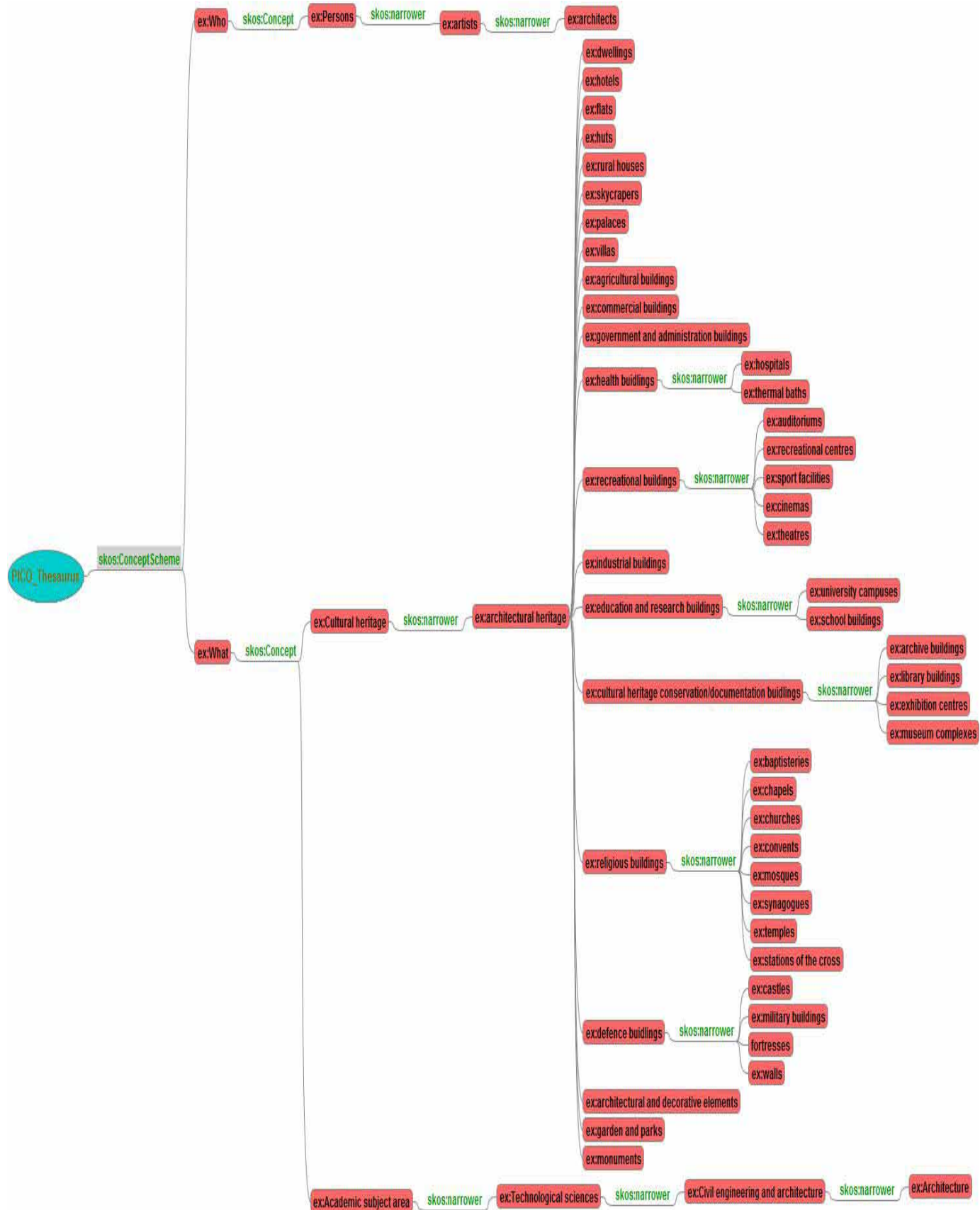
7.5. Schematic view of the MICHAEL Subject list for Architecture domain



7.6. Schematic view of the RMCA Thesaurus for Architecture domain



7.7.Schematic view of the PICO Thesaurus for Architecture domain



7.8. Schematic view of the ATHENA Thesaurus

