Semantic Interoperability Centre Europe

Guidelines and Good Practices for Taxonomies

Issue: 1.3

Date: 2009-10-26

Authors: Fraunhofer ISST /]init[

Document Change History

Date	Version	Author	Change Details
2009-06-29	0.5	HA, AB, JE	Internal Draft
2009-07-10	0.9	HA, AB, JE	Final Draft
2009-07-15	0.9.3	HA, AB, JE	Section 3.2.5 added, QA
2009-07-15	0.9.4	RFS	Comments
2009-07-23	1.0	HA, AB, JE	Final Version
2009-09-03	1.1	HA, AB, JE	Including the results of the meeting from August 24th
2009-09-07	1.1.1	HA, AB, JE	Small Corrections
2009-10-12	1.2	HA, AB, JE	Changes according to the review by Aldo Laudi
2009-10-26	1.3	HA, AB, JE	Final Version

In memory of our dear colleague
Helmut Adametz
(1956-2009)

Table of Contents

Table of Figures

PREFACE

About SEMIC.EU

SEMIC.EU (Semantic Interoperability Centre Europe) is an EU Project to support the data exchange for pan-European e-Government services. Its goal is to create a repository for interoperability assets that can be used by e-Government projects and their stakeholders. SEMIC.EU offers the following services for the public sector in Europe:

1 SEMIC.EU will provide access to interoperability assets that have been developed in previous governmental projects.

2 A clearing process will safeguard certain rules and standards to assure the quality of published assets.

3 Community features will be available on the platform, e.g. a forum to discuss best practices for the use of assets.

4 SEMIC.EU will invite stakeholders to seminars and workshops that are related to its activities.

5 SEMIC.EU offers coaching services for the creation and/or reuse of interoperability assets.

More information on SEMIC.EU can be found at: http://www.semic.eu.

SEMIC.EU is an action of IDABC. Contracted technical service providers for the project are:]init[(main contractor), Fraunhofer ISST, GEFEG, and France Telecom R&D.

About IDABC

IDABC stands for Interoperable Delivery of European e-Government Services to public Administrations, Business, and Citizens. It takes advantage of the opportunities offered by information and communication technologies to encourage and support the delivery of cross-border public-sector services to citizens and enterprises in Europe and to improve efficiency and collaboration among European public administrations.

The programme also provides financing to projects addressing European policy requirements, thus improving cooperation among administrations across Europe. National public-sector policy makers are represented in the IDABC programme's management committee and in many expert groups. This makes of the programme a unique forum for the coordination of national e-Government policies.

http://ec.europa.eu/idabc

Conventions

The type styles shown below are used in this document to emphasize parts of the text.

Times New Roman – 11 pt.: Standard body text

Times New Roman – 11 pt. Italic: Citations

The requirements level indicators are fully aligned to “*RFC2119 - Key words for use in RFCs to Indicate Requirement Levels*” and are used as follows:

MUST means that this policy element or requirement is to be fulfilled without exception.

SHOULD indicates an optional policy element / requirement that may be fulfilled if desired.

MANAGEMENT SUMMARY

The objective of SEMIC.EU is fostering semantic interoperability in the context of pan-European data exchange among public administrations. A clear and precise description and structuring of the information in the given domain are prerequisites for a common understanding of the information exchanged among different partners. Taxonomies and other types of controlled vocabularies are the preferred means to achieve such a common understanding by specifying the terms of the domain, disambiguating them from each other, controlling synonyms, and structuring the domain via term relationships.

In a pan-European context, partners from different Member States are typically involved. They usually use different taxonomies, and possibly ones in different languages. To reach a common understanding based on such different taxonomies, mappings and mediations are necessary.

This document focuses on taxonomies because they contain all the features that are essential when constructing even more elaborated types of controlled vocabularies. In addition, taxonomies are the type of controlled vocabulary most often used in practice. For the sake of completeness, however, some additional topics are also dealt with, e.g. associative relationships and terminological expressions.

The whole life cycle of taxonomies is treated, including the technical, methodological, conceptual, as well as organisational aspects. In order to build good taxonomies, the following steps are indispensable:

- implementing a structured process for the development and maintenance of taxonomies analogous to proven process models in the discipline of software engineering
- considering the reuse of one ore more existing taxonomies, e.g. from the SEMIC.EU platform or seeking for similar projects on SEMIC.EU to launch a collaboration and a related quality process
- choosing the appropriate type of taxonomy with regard to flexibility and expressivity.

Generally, taxonomies have to be implemented in a machine readable language. In practice, pan-European partners use a diversity of taxonomy languages, which necessitates syntactic integration. When integrating taxonomies syntactically, it is essential to

- rely on existing standards based on the paradigms prevailing in the relevant communities
- use a systematic translation mechanism between syntactically heterogeneous taxonomy languages based on a reference architecture.

In the case of semantically heterogeneous taxonomies, e.g. when a domain is ordered according to two different classifications, semantic integration has to be performed. This document treats

- proven matching techniques that serve to detect heterogeneous parts within taxonomies
- mappings for resolving heterogeneity when merging existing taxonomies as well as when mediating among decentralised taxonomies.

1. INTRODUCTION

The fundamental objective of SEMIC.EU is to foster semantic interoperability. The important layers of interoperability in the context of SEMIC.EU are shown in Figure 1.

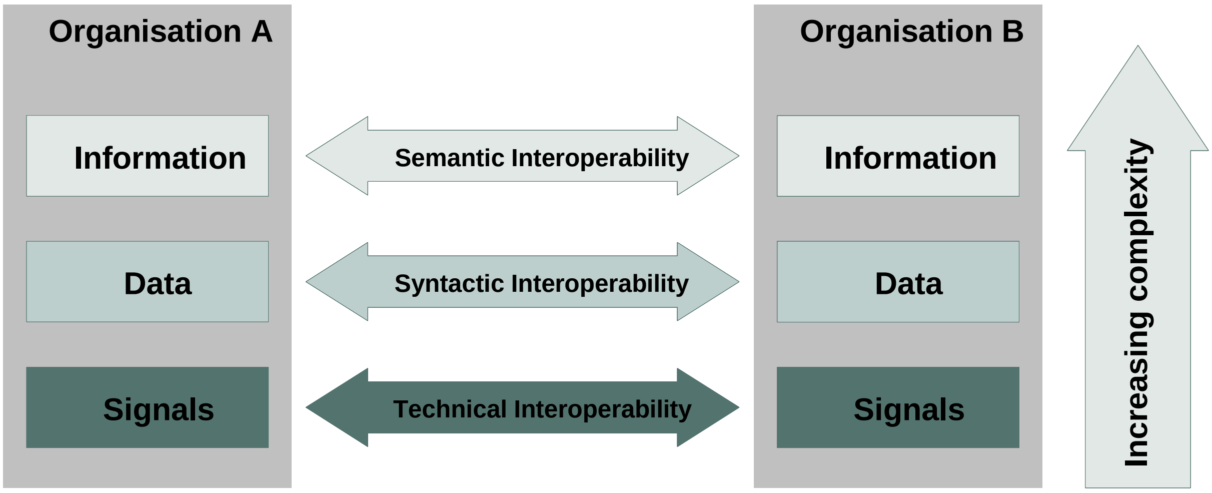


Figure 1: Layers of Interoperability

Technical interoperability, i.e. the interoperability on the protocol level, and organisational interoperability (not depicted in the above figure), i.e. the interoperability between communication partners related to underlying business processes and related agreements, are out of the scope of SEMIC.EU. Syntactic interoperability is considered, as long as it is related to semantic interoperability.

Semantic interoperability is mainly based on a common, unambiguous understanding of the information exchanged between different partners, mainly of the concepts used within this information. The means to achieve this is provided by taxonomies. Good taxonomies achieve semantic operability by

- Describing all relevant terms within the domain of the information exchange
- Completely eliminating ambiguity
- Controlling all synonyms within the domain
- Establishing appropriate relationships between terms [ANS05].

The partners involved in communication may base the understanding of the information they exchange on a common taxonomy, or they may use different taxonomies. If – the usual case in the context of SEMIC.EU – partners in different Member States are involved, they may use multilingual taxonomies, or they may use different taxonomies in different languages. When different taxonomies are used, translations, mappings, or mediations between these taxonomies are necessary.

Taxonomies can be means of supporting interoperability; they can also be the subject of data exchange. Nevertheless, implementing these taxonomies in different contexts implies applying the same methods and techniques as when they are used as a means of support for interoperability.

These guidelines are focused on taxonomies (plus additional features from thesauri), as this type of controlled vocabulary[1] comprises the features that build the core of more advanced vocabularies and accounts for most of the controlled vocabularies in practical use.

1.1 The Purpose of this Document

These guidelines present technical, conceptual, and organisational means and good practices to identify and tackle all kinds of issues arising when developing, maintaining, reusing, and integrating taxonomies. More precisely, with regard to the development and maintenance of taxonomies, the reader will come to understand how to structure the processes and also learn about concrete principles and techniques to apply in each of the steps involved.

With respect to the integration of different taxonomies, a description is given of the factors that have to be considered when using and integrating different taxonomy languages syntactically. Moreover, issues of heterogeneity and semantic integration of taxonomies are discussed with a view to the applicability of related methods and techniques.

This document is based on actual standards and guidelines related to the development and maintenance of taxonomies. In addition, it draws on the current state of research in the field of knowledge representation up to recent results from ontology mapping and alignment, as far as they apply to the integration of taxonomies. It has to be noted that some fields, e.g. (semi-)automatic techniques for identification of terms and relations or automatic matching of concepts are only marginally - or not at all - considered in the standards and guidelines. This is the case despite the fact that they play an important role in current research.

The scope of the problems addressed is illustrated by a number of typical usage scenarios and questions derived from them. In this sense, the presentation abstracts from concrete projects, but the findings are nonetheless intended to be practicable.

1.2 The Structure of this Document

This document is structured as follows:

- In section two, a detailed definition of taxonomies is provided.
- In section three, typical usage scenarios and related questions are introduced. These scenarios cover the problems typically encountered when dealing with taxonomies.
- In section four, a structured process for developing and maintaining taxonomies is presented including principles and practises for building good taxonomies.
- Section five deals with the syntactic integration of taxonomies represented in different languages. Different expressivity levels corresponding to different standardisations and an integration architecture facilitating the exchange of taxonomies written in different languages are described.
- Section six treats various kinds of correspondences between semantically heterogeneous taxonomies. In addition, techniques for detecting and handling this heterogeneity are presented. Explanations are given as to how these techniques can be applied for merging and translating taxonomies, as well as building mediators between them.
- The main issues of this document are compiled in the conclusions.
- The appendix summarises the rules, techniques, and good practices applicable within the life cycle of taxonomies.

[1] For an overview of all types of controlled vocabularies see [FI09].

2. BASIC PRINCIPLES

The general purpose of taxonomies is the organisation of information. The terms from a taxonomy are used to precisely and unambiguously describe (content) objects usually within a specific domain. According to [ANS05], taxonomies[2] serve five purposes:

1 Provide a vocabulary that can be used for indexing and retrieval

2 Promote uniformity in term format and in the assignment of terms

3 Indicate semantic relationships among terms

4 Provide consistent hierarchies for navigation to help users locate desired content objects

5 Serve as a search aid in locating content objects

In the course of this section, the constituents of taxonomies are presented.

2.1 Terms – Scope and Form

2.1.1 Scope of Terms

The basis of taxonomies consists of terms representing concepts, i.e. units *"of thought, formed by combining some or all of the characteristics of a concrete or abstract, real or imaginary object …"* [ANS05]. Concepts may range from rather abstract categories to instances, i.e. unique objects.

The following examples list possible categories and related terms:

Things and their physical parts

Birds
Documents
Monuments
Mountain regions

Disciplines or subject fields

Archaeology
Biology
Chemistry

Places

Australia
South Kensington
Sri Lanka

In the last of the examples, the terms designate unique objects/instances.

The terms have to be unambiguous within the domain of the taxonomy in which they are used. Therefore, homographs, i.e. terms with the same spelling but representing different concepts within one taxonomy, have to be disambiguated by qualifiers:

Cranes (bird)
Cranes (lifting equipment)

See section 2.2.1 regarding the treatment of synonymy within a taxonomy.

[2] In fact, [ANS05] states these purposes for controlled vocabularies.

The meaning of a term can be further determined by a scope note. It serves to restrict or expand the application of a term, distinguish between terms with overlapping meaning in natural language, or provide other advice on term usage [ANS05] [BS05]. When references to other terms occur in a scope note, reciprocal scope notes for each of these terms should be provided.

A scope note does not need to be a full definition. When needed, a definition can be added. To track the development of a term over time, a history note can be added.

2.1.2 Form of Terms

Issues to be considered concerning the form of a term are grammatical form, capitalisation, punctuation and special characters, singular or plural forms, the use of compound terms (e.g. rules for splitting and rules when not to split compound terms), etc. These issues are not treated within this document; they are comprehensively treated in [ANS05] and [BS05]. Though both of these standards are aimed at the monolingual cases of American and British English, respectively, the guidelines they present can be applied or easily adapted to other languages. A less comprehensive treatment for the multilingual case is provided in [AW80].

2.2 Relationships

Terms within a taxonomy don't exist in isolation from each other; rather, they are related by different semantic relations expressed by equivalent, hierarchical, and associative relationships. Though taxonomies and classifications don't contain associative relationships, nevertheless, for the sake of completeness, they are addressed within this document.

2.2.1 Equivalence Relationships

There is an equivalence relationship between terms when they express the same concept. There are different types of equivalence relationships:[3]

- Synonyms
- Lexical variants
- Near-synonyms

SYNONYMS

Synonyms are terms regarded as having the same meaning in a wide range of contexts. Examples of (different types) of synonyms (excerpt from [ANS05]) are:

Synonyms of different linguistic origin

cats / felines
freedom / liberty
sodium / natrium

Dialectical variants

elevators / lifts
subways / undergrounds

[3] [ANS05] treats the two further types "Generic Posting" and "Cross-references to Elements of Compound Terms", which for the sake of simplicity are not considered in this document.

LEXICAL VARIANTS

Whereas synonyms are different terms representing the same concept, lexical variants are different word forms for the same term or expression:

ground water / ground-water / groundwater
mice / mouse

NEAR-SYNONYMS

Near synonyms are terms that actually do not have the same semantics, but are treated as equivalent in the context of the taxonomy. The following example shows three common types of near-synonyms.

sea water / salt water	[variant terms]
meteors / meteorites / meteoroids	[point on a continuum]
smoothness / roughness	[antonyms]

Within a taxonomy, one of the terms representing the same concept has to be distinguished as the preferred term among these synonyms.

2.2.2 Hierarchical Relationships

Hierarchical relationships are the primary feature of taxonomies and classifications[4], and actually "*have a central role in almost all domain ontologies …*" [SCK+05]. Hierarchical relationships are based on levels of superordination and subordination, where a (broader) superordinate or parent term represents a class or a whole, and a (narrower) subordinate or child term represents a subclass or a part of the whole. Every term in a taxonomy has to have a hierarchical relationship to at least one other term within the taxonomy.

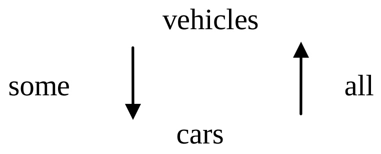
Hierarchical relationships are not a homogeneous principle, but comprise three logically different and mutually exclusive types of subsumptions [ANS05]:

- The generic relationship
- The whole-part relationship
- The instance relationship

GENERIC RELATIONSHIP

This generic relationship is characterised by a subset relationship between the extension of the narrower term and the extension of the broader term, i.e. each object captured by the broader term is also an object captured by the narrower term. The generic relationship can always be expressed as an "is a" relationship, i.e. "[narrower term] is a [broader term]". The validity of the relationship can be tested by an "all-and-some" test [ANS05]:

[4] This document does not distinguish between taxonomies and classifications, although classifications - in contrast to other controlled vocabularies - don't have preferred terms. Instead, each concept is unambiguously related to a unique notation, e.g. a decimal number.



Instance Relationship

The instance relationship is also characterised by an “is a” relationship between the narrower term and the broader term. However, in contrast to the generic relationship, the narrower term in an instance relationship does not express a set of objects, but a specific single object:

Mountain regions

Alps
Himalayas

Whole-Part Relationship

The whole-part relationship “*refers to the relation between a concept/entity and its constituent parts*” [KN06]. This relationship can be characterised by an “is part of” relationship between the narrower term and the broader term:

EU

Spain
Italy

Nevertheless, the whole-part relationship comprises different types. According to [SS08], segmentation of structures can be distinguished from segmentation independent of structures.

Segmentation of structures	
Geographical unit / subunit	Canada / Ontario
Collection / element	Wood / tree
Organisation / unit	Company / department
Complex / component	House / roof
Event / segment	Show / scene
Segmentation independent of structures	
Whole / piece	Bread / slice (of bread)
Activity / phase	Procurement / payment
Object / part	Seat / wooden parts (of seat)
Mass / portion	Meter / decimeter

Table 1: Specific Whole-Part Relationships [SS08]

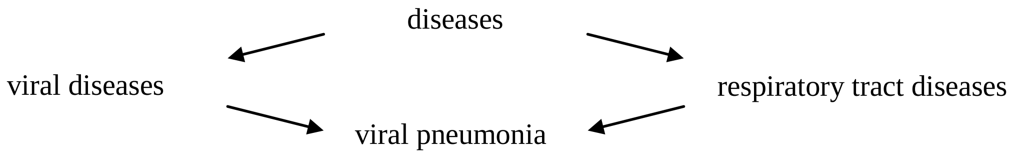
It has to be noted that the whole-part relationship is not transitive when different types of whole-part relationships are mixed [WCH87]:

> *"Simpson's finger is part of Simpson*
>
> *Simpson is part of the Philosophy Department*
>
> *? Simpson's finger is part of the Philosophy Department"*

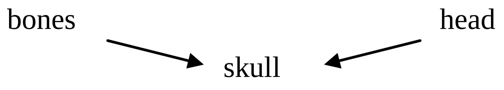
Even within the types described above, different variants may occur that impede transitivity, e.g. membership of an organisation within other organisations versus membership of single individuals within an organisation [Joh04].

2.2.3 Polyhierarchical Relationships

A taxonomy is monohierarchical when each term except the top term has one and only one broader term, otherwise, the taxonomy is polyhierarchical. In this case, the broader terms are not disjunctive, but overlapping. An example of a large polyhierarchical taxonomy is the Medical Subject Headings Thesaurus [MSH08], from which the following example, based on generic relationships, is taken:



Polyhierarchical relationships may be based on different relationships. In the following example (according to [ANS05]), bones and skull are in a generic relationship, whereas head and skull are in a whole-part relationship:



2.2.4 Faceted Classification

When and only when all the objects within a domain can be described by the same attributes [SS08], a faceted classification of these objects may be appropriate. [Bro04] provides a simple example that illustrates this approach. A classification of socks in hierarchical form is given by:

Grey socks

Grey wool socks

Grey wool work socks
Grey wool hiking socks

Grey wool ankle socks for hiking
Grey wool knee socks for hiking

Grey spotted wool knee socks for hiking

In a corresponding faceted classification, objects would be classified according to the attributes and dimensions, respectively:

Colour	Pattern	Material	Function	Length
Black	Plain	Wool	Work	Ankle
Grey	Striped	Polyester	Evening	Calf
Brown	Spotted	Cotton	Hiking	Knee

Whereas the faceted classification shown in this example with 15 terms in five dimensions provides 273 possible combinations, an equivalent hierarchical taxonomy would have to provide a separate term for each of these combinations. Thus, faceted classifications are much smaller than corresponding hierarchical taxonomies. It has to be noted that, in general, some combinations may be meaningless, e.g. polyester socks may not be suited for hiking [TSCA02].

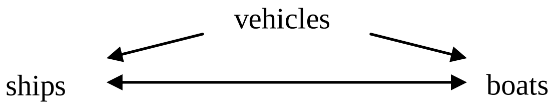
Whereas in the example for each dimension the terms are provided as a flat list, in the general case, for each dimension a separate hierarchical taxonomy could be provided.

2.2.5 Associative Relationships

Associative relationships are the additional feature of thesauri (and ontologies) that distinguishes them from taxonomies. Associative relationships indicate associations between terms that are neither equivalent nor hierarchical, yet they are semantically related in some sense. The associative relationship of cells and cytology in the example below is based on the fact that cells constitute a necessary part of the definition of cytology:

cells ↔ cytology

Associative relationships may occur between terms belonging to the same hierarchy (see example) or between terms belonging to different hierarchies:



Associative relationships in taxonomies often have no specified type. Such a relationship may be interpreted as “see also” and usually is symmetrical. Nevertheless, associative relationships may be assigned to indicate specific types of relations that, in general, are not symmetrical. The following table lists such types of relations and related examples.

Cause / Effect	Accident / injury
Action / Product	Writing / publication
Action / Property	Communication / communication skills
Action / Target	Teaching / student
Raw material / Product	Grapes / wine
Discipline or Field / Object or Practitioner	Neonatology / infant

Table 2: Associative Relationship Types (Excerpt from [ANS05])

Associative relationships may be named to indicate the type of relationship they express. In the following example, an excerpt of named associative relationships from the AGROVOC Thesaurus [AGR09] is listed:

causative

controls
preventedBy
actsUpon
causedBy
performedBy
producedBy

instrumental

performedByMeansOf
growthEnvironmentFor
growsIn

2.3 Ontology Features

Ontologies enhance the features of taxonomies by the following constructs:

- The thesaurus constructs described in section 2.2.5
- Constructs for defining schemata like attributes for defining properties of concepts including related data types, e.g. an attribute “age” with related data type “integer”.
- Moreover, full-fledged ontologies introduce constructs that allow for stating complex logical expressions over concepts or defining rules over concepts.

On the one hand, ontologies are more expressive than taxonomies, but, on the other hand, ontologies are much harder to construct, maintain, and understand than taxonomies (see also section 4.3.5, example “Gene Ontology”). Currently, ontologies are much less widespread than taxonomies in practice.

Ontologies and related constructs will not be treated in detail in these guidelines; for an introduction to ontologies see [FI09].

2.4 Multilingual Taxonomies

In the context of SEMIC.EU, multilingual taxonomies and classifications play an important role. In this document, a multilingual taxonomy is assumed to be symmetrical, i.e. “*all different language versions of a multilingual thesaurus have to be identical and symmetrical; each preferred term must have one and only one equivalent term in every language and be related in the same way to other preferred terms in the given language (a symmetrical thesaurus). [...] The number of non-preferred terms can be different”* [IFLA09]. Such a symmetrical taxonomy (thesaurus) corresponds to the unity model presented in section 6.4.1.

Concerning the equivalence of preferred terms in different languages, a distinction can be made between exact equivalence on the one hand, and inexact or near equivalence on the other hand (see section 6.4.1, where the different types of equivalence are described in detail). Whereas exact equivalence corresponds to true synonymy in the monolingual case, inexact or near equivalence corresponds to near-synonymy in the monolingual case (see section 2.2.1), which is treated as exact equivalence.

When a suitable term for a concept is missing from one of the languages in a multilingual taxonomy, there are two solutions. Either a loan term from another language can be used [BS08]:

German	English
Schadenfreude Teenager	Schadenfreude Teenager

or an artificial so-called “coined term” can be defined:

English	French
gender mainstreaming	Intégration de la dimension de genre

A second type of multilingual taxonomy is the non-symmetrical taxonomy, where the number of preferred terms in each language does not need to be the same, and where the relationships between preferred terms can differ for the different languages [IFLA09]. Such a multilingual taxonomy can be considered as being not one taxonomy, but consisting of a number of heterogeneous taxonomies, one of each language. According to [BS08], in this case, the different language versions have to be mapped to each other applying the mapping techniques presented in section 6.4.1.

This section introduced the constituents and the structure of a taxonomy, i.e. the terms with additional notes attached to them and the different types of relationships between these terms. Moreover, the different types were presented with regard to hierarchy, i.e. monohierarchy, polyhierarchy, and faceted classification. Finally, some aspects of multilingual taxonomies were introduced. The next section will outline various scenarios and related questions that occur when developing and maintaining taxonomies, as well as scenarios addressing the syntactic and semantic integration of different taxonomies.

3. Usage Scenarios

In this section, some typical scenarios will be presented that may occur within the life cycle of taxonomies and that span the scope of the principles, techniques, and good practices presented in the subsequent sections of this document. The first and most prominent scenario is the reuse of an existing taxonomy, e.g. a taxonomy found on the SEMIC.EU platform, as reuse always entails the potential of saving costs and increasing quality by reusing mature and tested solutions.

Each presentation consists of a general description of the scenario and questions that arise in the context of the scenario. For each question, a reference to the section where it is treated is added.

3.1 Reuse of an Existing Taxonomy

Rather than developing a new taxonomy from scratch, it is highly recommended to find one ore more suitable existing ones that can be reused, e.g. on the SEMIC.EU platform that serves as the European platform for providing reuse candidates or at least for informing about similar developments in progress or being in the planning phase.

The following questions arise in the context of this scenario:

- Which cost factors are implied by the reuse of a taxonomy (see section 4.2)?
- What are possible benefits when reusing a taxonomy (see section 4.2)?
- Which parameters do have an impact on the reusability of a taxonomy (see section 4.2)?
- What are positive factors for a possible reuse (see section 4.2)?
- What has to be considered when more than one taxonomy shall be reused (see sections 4.2 and 6.4.2)?
- How can the decision about reusing one ore more existing taxonomies be supported (see section 4.2)?

3.2 Development of a New Taxonomy

This scenario consists of the development of a completely new taxonomy from scratch, after failing to find a suitable taxonomy for reuse. Thus, the development is not impeded by requirements and restrictions, e.g. by having to comply with other taxonomies.

It is highly recommended to start the development with registering the project on the SEMIC.EU platform in order to inform the community and possibly find collaboration partners. A forum thread could be initialised including a call for comments. In the course of the project, it is recommended to continuously publish new versions of the taxonomy and to achieve passing SEMIC.EU's maturity or even conformance process in order to improve the quality of the taxonomy.

The following questions arise in the context of this scenario:

- Which principal requirements have to be considered for the development of a taxonomy? (see section 4.1)
- How should the objects/terms to be included in the taxonomy be identified? (see section 4.3.1)
- Which criteria should be used for classification of the objects of the taxonomy? (see section 4.3.2 to section 4.3.4)
- Are there different independent criteria to classify the objects of the taxonomy, e.g. branch, topic, geographic location, etc.? (see section 4.3.4)

- What is the appropriate granularity for the taxonomy, e.g. how many child terms should a parent term have? (see sections 4.3.1 and 4.3.2)
- Which kinds of relationships among the objects should be chosen for inclusion in the taxonomy? (see sections 4.3.1 to 4.3.5)
- Should the meaning of the terms be disjunctive, or should overlapping of the meaning of the terms be allowed? (see section 4.3.3)

3.3 Maintenance of a Taxonomy

A taxonomy has been released and is in operation, possibly in different environments and contexts. Usually, errors will occur that have to be fixed. The taxonomy may not be complete, i.e. users may not be able to find particular terms, both general and specific. Furthermore, there may be terms that are not used at all. Additional requirements may occur concerning a use of the taxonomy in an additional application.

The following questions arise in the context of this scenario:

- What kinds of changes do occur during maintenance? (see section 4.6)
- What are the reasons for changes? (see section 4.6)
- When are changes necessary or advisable? (see section 4.6)
- Who has the authority to decide about changes? (see section 4.6)
- What are possible implications of modifying the taxonomy? (see section 4.6)
- Are changes to be propagated immediately? (see section 4.6)
- How are the changes to a taxonomy to be disseminated? (see section 4.6)

3.4 Extension of a Taxonomy

Extensions of a taxonomy are usual activities within the development when reusing another taxonomy or in the maintenance process of a taxonomy due to new requirements. There are, however, cases where a taxonomy has to be realised as an extension of a taxonomy, i.e. the given taxonomy may be refined or enhanced, but not changed otherwise.

The following questions arise in the context of this scenario:

- How can a given taxonomy be enhanced on the technical level in contrast to enhancing a taxonomy in the usual maintenance process? (see section 4.9)
- Which implications do result from a modification of the superordinate taxonomy? (see section 4.9)
- How is an extended taxonomy to be maintained, i.e. which variations from the regular maintenance process may occur? (see section 4.9)
- Which organisational variations are necessary in comparison to the usual development and maintenance of a taxonomy? (see section 4.9)

3.5 Development of a Multilingual Taxonomy

In the context of SEMIC.EU, multilinguality is the usual case. Therefore, for partners and related applications that seek to exchange data on a pan-European level, it may be advisable to base communication on a multilingual taxonomy. As in the monolingual case, before starting an own development, it is highly recommended to find one ore more suitable existing ones on the SEMIC.EU platform that can be reused (see section 3.1). Possibly, although no suitable multilingual one could be found, there is a suitable monolingual one. In principle, the same questions as stated in section 3.1 apply.

In general, multilinguality implies or is accompanied by political and cultural differences that may have an impact on the multilingual taxonomy.

The following questions arise in the context of this scenario:

- Can a multilingual taxonomy be developed by enhancing a monolingual taxonomy? (see section 4.8)
- How is a multilingual taxonomy to be structured? (see section 2.4)
- How should a multilingual taxonomy be maintained, i.e. which variations from the regular maintenance may occur? (see section 4.8)
- Are there organisational impacts different from the monolingual case? (see section 4.7)

3.6 Syntactic Integration and Exchange of Taxonomy Languages

In this scenario, SEMIC.EU project partners seek to achieve the syntactic integration of different taxonomy languages by means of standards, which demands the use of efficient and adequate intermediary structures. Partners have to be aware of various expressivity levels and paradigms of taxonomy languages in order to correctly exchange related taxonomies.

The following questions arise in the context of this scenario:

- What are the drawbacks of using customised languages for integration? (see section 5.1)
- Which technological paradigms do stand behind taxonomy languages? (see section 5.2)
- Which core techniques are used for advanced taxonomy languages? (see section 5)
- How are intermediary structures used in the context of syntactic integration? (see section 5.4.2)
- What are the principal use cases for syntactic integration? (see section 5.4.2)

3.7 Integration of Heterogeneous Taxonomies

The integration of taxonomies that model the same domain under different views constitutes the present scenario. Techniques for detecting heterogeneity as well as methods to define correspondences between heterogeneous taxonomies have to be applied. The use of mappings for the process of merging, translating, and mediating taxonomies are part of this scenario.

The following questions arise in the context of this scenario:

- Which kinds of heterogeneity do exist? (see section 6.2)
- Which techniques do exist for matching heterogeneous taxonomies, and how are they related to taxonomy expressivity? (see section 6.3)
- What kinds of taxonomy mappings do exist? (see section 6.4.1)
- How are mappings used for merging, translation, and mediation of heterogeneous taxonomies? (see section 6.4.2)

3.8 Integration of Heterogeneous Taxonomies from Different Domains

Beside the integration of taxonomies modelling the same domain, the case has to be considered where taxonomies from different domains have to be integrated. This, for example, is the case when for a cross-domain application existing taxonomies from different domains shall be reused (see also section 3.1). If the taxonomies to be integrated have overlapping parts, i.e. overlapping sub-domains, the same questions as in the scenario described in section 3.7 apply.

The following questions arise in the context of this scenario:

- What are the interrelations between domains and heterogeneous taxonomies? (see section 6.1)
- How to interrelate taxonomies from different domains? (see section 6.1)
- Do the taxonomies to be integrated overlap in common sub-domains? (see section 6.4.1)
- Are the same names used for different concepts within different domains and related taxonomies, respectively (see 6.4.2)?

The present section outlined typical scenarios and their respective issues. The following sections will address principles, techniques, and solutions concerning the questions compiled above. These solutions apply to the life cycle of a single taxonomy as well as to problems when several taxonomies are involved. The next section will discuss issues of development, maintenance, and extension of taxonomies including the multilingual case.

4. Principles, Methods, and Good Practices for the Development of Taxonomies

Construction of a taxonomy is a *"time-consuming, labor-intensive process, especially if the domain to be covered is broad and the terminology in use is rich and complex"* [ANS05]. According to [SS08], large classifications are the result of decades of work. Thus, it is indispensable to develop a taxonomy in a systematic way. Analogous to software engineering, a systematic approach should be applied in the development of taxonomies in order to guarantee the quality of their clarity, correctness, and consistency. Several methodologies and guidelines for the development of controlled vocabularies including ontologies have emerged in recent years [FG97] [NM01a] [FG03] [ANS05] [BS05] [Cho06] [WB08]. Each of these methodologies structures the development into several steps and activities that do not need to be performed in a strict order, but which are considered part of an iterative process. The following steps are mainly adapted from [WB08]:

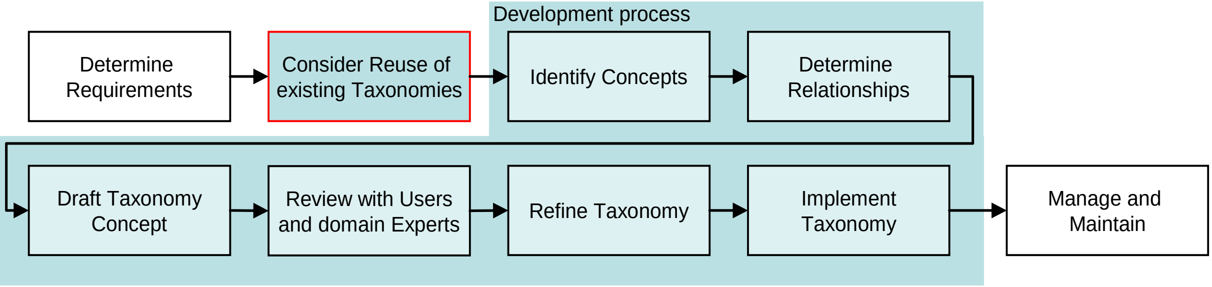


Figure 2: Development Process

The most important of these steps is the second one that demands to consider the reuse of existing taxonomies, as this is the best way to save costs and to increase quality by building on well-proven existing solutions as they are provided by the SEMIC.EU platform, in particular, when these solutions have passed SEMIC.EU's maturity or conformance process. But even when no existing solution can be found, similar projects may be found on the platform and a collaboration may be launched.

It has to be noted that reusing an existing taxonomy does not imply that the process steps following the second step become superfluous. This would only be the case when an existing taxonomy could be reused without any modifications. However, usually the following steps will be significantly less comprehensive than in the case of a development from scratch.

Each of the process steps should be accurately documented, e.g. which principles have been applied for constructing the taxonomy.

4.1 Determine Requirements

In this first step, several basic questions have to be answered:

- What is the purpose and objective of the taxonomy?
- What is the scope of the taxonomy?
- What is the intended use of the taxonomy?
- Who are the user groups of the taxonomy, e.g. experts and/or lay users?
- Who will maintain the taxonomy?

For each of these questions, possible future developments and enhancements, e.g. use in additional contexts and further user groups, have to be considered.

4.2 Consider Reusing Existing Taxonomies

Before creating a new taxonomy from scratch, it should always be investigated whether an existing one can be reused. As already mentioned above, SEMIC.EU as the primary pan-European resource for Sematic Interoperability Assets is the preferred location to look for reuse candidates.

The case studies presented in [BMT05] identify the following cost factors in regard to taxonomy reuse:

- Identification and the process of gaining familiarity with reuse candidates
- Translation of existing taxonomies, possibly in proprietary formats, into the target format
- Integration of various heterogeneous sources for reuse (applying the techniques described in section 5).

The possible benefits of reuse identified in [BMT05] are:

- Reduction of development costs
- Improvement of interoperability with applications using the same taxonomy.

In particular, reuse should be considered when

- There are well-established taxonomies available that are applied in contexts quite similar to the one in which the planned taxonomy is to be applied.
- One single reuse candidate covers most of the domain the target taxonomy has to cover, i.e. it is not necessary to merge different taxonomies.
- The identified reuse candidate is implemented in the same taxonomy language as the planned taxonomy is to be implemented in, or is at least implemented in a standard taxonomy language, not in a proprietary one (see section 5).

In [BM05] there is a cost model for ontology engineering. Based on established software engineering approaches, it is proposed that a (sub-)model be included for reuse. This model or future models serving the same purpose should be applied in order to conduct a cost/benefit analysis before deciding whether to reuse existing taxonomies or to build one from scratch.

Nevertheless, the quite recent source [AELP09] states for ontologies, but also applicable for taxonomies that *“even with so many ontologies now available to reuse, the tasks of finding, assessing and effectively making use of existing ontologies remain difficult”*, mainly due to *“the lack of support for ontology practitioners to find, assess and exploit existing ontologies.”*

As in most cases existing taxonomies cannot be used without modifications, adaptations, or extensions, even when taxonomies are to be reused, the following steps in the development process have to be taken, even though usually they will be less extensive.

4.3 Identify Concepts and Determine Relationships

The next two steps in the taxonomy development process are identifying concepts and determining relationships among them. These steps are the core activities in this process. Whereas some approaches propose performing these steps in succession, others propose performing them in parallel. Furthermore, top-down approaches are distinguished from bottom-up approaches.

The top-down approach starts with a definition of the most general concepts in the domain, followed by a specialisation of the concepts [UG96]. As sources for finding these concepts, studies, textbooks, reference texts, encyclopaedias, etc. about the domain may be used. The *“hierarchical structures and relationships are created as the work proceeds”* [ANS05]. The bottom-up approach starts with determining the specific terms. In the case of classifying documents, these terms can be extracted from the documents. As in the top-down approach, the hierarchy and the relationships are created in parallel.

In practice, in most cases, a combination of these approaches will be applied. Due to [NM01a], none of the three approaches, i.e. top-down, bottom-up, or a combination of the two, is inherently better than any of the others. The choice of the approach may depend on the available sources. When sources as reference texts and encyclopaedias contain a systematic, top-down approach to the domain including definitions of categories, the top-down approach may be preferred. When the main sources available are the content objects to be described, e.g. documents to be classified, the bottom-up approach may be preferred. According to [Ros78], the combination may be preferred since the concepts "in the middle" tend to be the more descriptive concepts in the domain.

[ANS05] subsumes these approaches as the committee approach and distinguishes them from the empirical approach, which subsumes the deductive and inductive method. The main difference between these methods is that in the deductive method, *"terms are extracted from content objects [...], but no attempt is made to control the vocabulary, nor to determine relationships between terms, until a sufficient number of terms have been collected"* [ANS05]. In contrast, in the inductive method, *"controlled vocabulary construction is regarded from the outset as a continuous operation"* [ANS05]. This is, in the deductive approach, the selection of terms and the construction of the taxonomy are performed sequentially, whereas in the inductive approach these tasks are performed in parallel. These approaches can also be combined, e.g. relationships developed inductively can later be examined from a deductive perspective [ANS05],. But [ANS05] does not characterise one of these methods, i.e. the deductive, the inductive, or the combined method as being superior for specific requirements or contexts, so the choice becomes a matter of personal preference.

In particular, when developing a complex taxonomy based on a huge number of text documents to be classified, the use of (semi-)automatic techniques should be employed to identify terms as well as relationships among them. Machine assistance, i.e. text mining tools, can be used to identify candidate terms by evaluating (frequency of) occurrence within relevant documents, and in particular, within content objects that are to be indexed by the taxonomy. The number of possible terms should be reduced by a stop list containing articles ("a/an" and "the"), prepositions ("to", from", conjunctions ("and", "yet"), etc. High-frequency terms are the most likely candidates for inclusion in the taxonomy. Relationships among terms may be derived from co-occurrences within relevant documents, applying mathematical and statistical methods such as rankings methods and cluster analysis [PC96] [SB05]. Moreover, text mining tools allow the definition of rules for identifying patterns within text that define relationships. For example, the pattern "A belongs to genus B" defines a hierarchical relationship between A and B. In general, such rules are domain specific and therefore have to be defined by domain experts.

Which terms should be included in a taxonomy and how the relationships should be determined is treated in the following subsections. Moreover, several types of taxonomies, e.g. monohierarchical, polyhierarchical, etc. are discussed concerning their appropriateness for the intended purpose.

4.3.1 Selection of Terms

Which terms should be included in a taxonomy is mainly determined by its domain and purpose. The selection of terms is governed by

- the vocabulary used within the domain as described in reference sources, textbooks, encyclopaedias, etc. about the domain
- the vocabulary that all of the user groups of the taxonomy are familiar with, in particular, terms that are used for information retrieval (see section 2, purpose 1).

Concerning the specificity of terms, according to [ANS05], the addition of highly specific terms is usually restricted to the core area of the subject field. In a taxonomy for the ceramics industry, "porcelain", "bone china", and "crockery" might represent different concepts, whereas for more general use, they could be treated as near-synonyms for the same concept [Dex01]. In general, near-synonyms should not occur within the core area of a domain. These concepts should be *"individually defined and retained"* [ANS05], as the differences among near-synonyms within the core area does matter.

True synonyms that are common within the addressed user groups should be included in the taxonomy.

Furthermore, in regard to the specificity and granularity of terms, there is a trade-off between the taxonomy's economy and its information content [Kom92]. If the terms are rather general, there will be relatively few terms which in turn increases the economy of the taxonomy. In this case, though, there will also be fewer characteristics shared by the objects captured by a general term, making it harder to identify these objects as identical and implying a decreased information content of the taxonomy. On the other hand, many specific concepts decrease the economy of the taxonomy. However, as there will be many characteristics associated to a specific term, the objects captured by this term can be identified more easily as being identical, implying an increased information content of the taxonomy.

Whether the terms within a taxonomy have to be disjunctive or whether terms with an overlapping meaning can be included, is discussed in section 4.3.3.

Scope notes, definitions, and history notes should be provided for each term included in the taxonomy as far as necessary. In particular, a definition is considered necessary when it cannot be assumed that a specific term will be understood by all of the user groups, e.g. when the German term "Ordnungswidrigkeit" is used as a loan term (see section 2.4) in a non-German taxonomy.

The terms included in a taxonomy must always comply with the rules for the form of terms as described in [ANS05] (see section 2.1.2 and section 2, purpose 2).

For basic applications and domains, the result of this step may be a glossary or code list respectively. Neither case, however, would include terms with differing levels of specificity, e.g. a list of the cities and towns within a state. But even for quite basic requirements, e.g. the requirement for grouping the cities and towns with regard to federal states, terms of different levels of specificity have to be considered in the term-selection process, and then the next step has to be performed, i.e. the terms have to be classified according to a hierarchy.

4.3.2 Determination of Hierarchy

The next section covers purposes 3 to 5 of section 2, namely indicate semantic relationships among terms, provide consistent hierarchies, and serve as a search aid for content objects.

Generic Relationship

How the hierarchy of a taxonomy is determined depends mainly on the generic relationship. In general, for each level of the hierarchy, each narrower term inherits the characteristics of its broader term and is distinguished from its sibling terms with regard to an additional property, as can be seen in the famous Tree of Porphyry [Bar03] (see Figure 3).

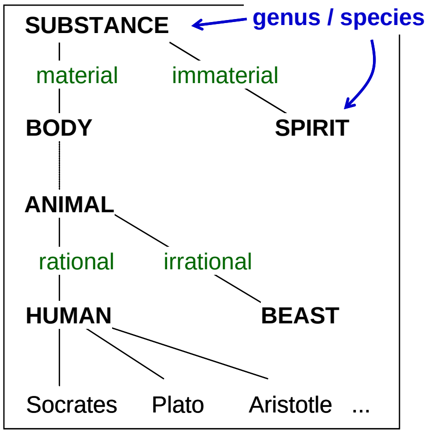


Figure 3: Tree of Porphyry

As the following simple example given by [HP05] illustrates, there is generally no natural or best classification of a set of objects, i.e. how to determine broader and narrower terms and how to choose the property by which siblings are distinguished. Supposed, the concepts to be classified are three figures, a black and a white square, and a black triangle.

■ ▲ □

It is not obvious how to classify the three figures. They may be classified according to their form or to their colour. But *"whether squares should form a class with triangles are excluded or whether black figures should form a class while white figures are excluded"* [HP05] depends only on the purpose of the classification.

Moreover, according to [Wit53], some concepts are not characterised by common properties, as is illustrated by means of games. There are board games, card games, and ball games that obviously are all games. Thus, for each of these kinds of games, a hierarchical relationship with "games" as a broader term could be defined. Some games have in common that there are winners and losers. Some games, but not all, are entertaining. Some require fortune and luck. Hence, games cannot be characterised by a common set of attributes and related properties. Even a disjunction of properties may characterise a concept. [Wit53] calls this kind of relation "family resemblance".

There are, however, constraints on the properties that can be used to distinguish between different concepts. For example, the height and shape of a plant is significantly determined by its habitat, i.e. the place where it grows and the environmental conditions related to it. Therefore, growth and shape are not suited to distinguish among different species, as instances of the same species may be different in height and shape.

Whole-Part Relationship

For taxonomies that use the whole-part relationship, for information retrieval purposes, transitivity should be granted [SS08]. A common search strategy is to add the broader term of a term into a search query when the original query does not deliver appropriate results. But adding further broader terms more than one level above only yields meaningful results when transitivity is given. Otherwise, there is no semantic relationship to the original search term (e.g. between Simpson's finger and the Philosophy Department in section 2.2.2).

GENERAL GOOD PRACTICES

Further general good practices or rules of thumb for the determination of the hierarchy of a taxonomy are the following [NM01a]:

- All the siblings in a hierarchy should have the same level of generality, as this corresponds to users' intuitive expectations.
- If a class has only one direct subclass, the taxonomy is not complete or modelling may be not appropriate.
  If a class has only one direct subclass, either the class or the subclass is superfluous, as either the class or the subclass doesn't add information to the taxonomy.
- If there are more than a dozen subclasses for a given class, adding intermediate categories should be considered.
  [NM01a] does not provide a justification of this rule but states that many well-structured ontologies have between two and a dozen direct subclasses. It is stated, however, that, *"if no natural classes exist to group concepts [...], there is no need to create artificial classes"*, as *"the ontology is a reflection of the real world."*

EXAMPLE ICD-10

The International Statistical Classification of Diseases (ICD-10) is a monohierarchical classification. The classification of the diseases is based on the principles of aetiology, i.e. the science of the cause of diseases, not according to their localisation. Thus, pneumonia is a narrower term of inflammation, not of lung, and other inflammations are siblings of pneumonia.

EXAMPLES IN BIOLOGY

The development of taxonomies in the field of biology over time stresses the importance of the role of the discrimination principle between the terms within a taxonomy. Whereas the famous Linnean Taxonomy is based on the morphology of plants or animals, i.e. shared physical characteristics, later taxonomies are based on cladistics, i.e. the evolutionary relationships among organisms, resulting in significantly different classifications. More recent approaches are based on DNS sequences, again implying different classifications.

4.3.3 Monohierarchy versus Polyhierarchy

According to [RM06], polyhierarchy is unavoidable when dealing with large information systems. In regard to expressivity and flexibility, polyhierarchical taxonomies are superior to monohierarchical ones, as the following example shows [Wol97]. In a museum, because of its form, an ancient tool may be classified as a hammer via an instance relationship. A further classification according to its function, e.g. "for shoeing horses", may be postponed until the tool has been thoroughly investigated. When this investigation has taken place, a further relationship to a broader term indicating the function may be added. Thus, a complete classification of the term can occur in several steps, which is a more flexible solution compared to a monohierarchical taxonomy. Moreover, browsing in polyhierarchies provides several paths to a content object and searching based on appropriate broader terms provides a greater number of possible search items.

On the other hand, statistical evaluations are impeded by polyhierachy since a content object may be counted several times. Most classifications, e.g. ICD 10 and NACE (Nomenclature général des activités économiques dans les Communautés Européens), prohibit polyhierarchy. In addition, polyhierarchical taxonomies tend to become difficult to understand, in particular, when every possible classification or distinction of terms is implemented in parallel.

Before implementing polyhierarchy, the use of a faceted classification should be considered (see sections 2.2.4 and 4.3.4).

EXAMPLE MESH

In the polyhierarchical Medical Subject Headings Thesaurus [MSH08] (see also section 2.2.3), diseases are classified by its elicitor as well as by its location. Therefore, the term "viral pneumonia" has the broader terms "viral diseases" and "respiratory tract diseases". This polyhierarchy could be resolved by a faceted classification with dimension "elicitor" and "location" and describing "viral pneumonia" by the terms "viruses" and "respiratory tract" (see next section).

4.3.4 Faceted Classification

When the condition for a faceted classification is fulfilled, i.e. when all objects within a domain can be described by the same attributes and dimension, a faceted classification should be implemented. It has to be noted that objects merely characterised by family resemblance (see section 4.3.2) do not fulfil this condition.

Faceted classification has the following advantages:

- A faceted classification is much smaller than a corresponding mono- or polyhierarchical one (see section 2.2.4) and therefore easier to construct and maintain.
- Faceted classification allows users to search or browse for resources with greater flexibility, as they can search for a resource from different angles [UJ07].

According to [ANS05], faceted classification is particularly useful for:

- New and emerging fields with incomplete domain knowledge
- Interdisciplinary areas with more than one perspective in order to look at content objects or those with a need for a combination of concepts
- Taxonomies with multiple hierarchies but unclear boundaries
- The classification of electronic documents and content objects where location and collocation are not of importance

However, due to [SS08], it is nearly impossible to change or modify the facets of a faceted classification during its life cycle.

EXAMPLE EPRACTICE

On the ePractice portal [EPR09], European eGovernment cases are classified according to the dimensions "Countries", "Domain", "Sector", "Status", "Type of Initiative" and "Regular Case Awards". These dimensions function as filters when searching for cases.

4.3.5 Adding of Associative Relationships

There are terms that are semantically or conceptually associated (see section 2.2.5), but neither in an equivalence nor hierarchical relationship. According to [ANS05], whether the association should be made explicit by an associative relationship depends on the extent of the association. Whether the extent justifies an associative relationship is determined by the purpose of the taxonomy. For instance, in a taxonomy containing products and related branches, unless the relation is expressed hierarchically, the relations should be made explicit by associative relationships.

Associative relationships enhance the usefulness and expressivity of a taxonomy by suggesting additional (associated) terms for use in indexing and retrieval [ANS05] and of named relationships in particular.

On the other hand, moving a taxonomy to a thesaurus by the use of associative relationship enhances its complexity (see example below). Moreover, associative relationships may impede interoperability among different thesauri. In particular, semantically different, i.e. not equivalent, named relationships cannot be mapped on each other.

Thus, the decision whether (named) associative relationships should be added or not depends on the indispensability of the semantic relation in the given application context.

EXAMPLE GENE ONTOLOGY

The famous Gene Ontology [GEN09] is a description of molecular functions, biological processes, and cellular localisations, maintained by an international consortium [LN07]. Despite the designation “ontology”, it is a taxonomy based on the generic and the whole-part relationship. On the one hand, this decision, i.e. the restriction to these basic relationships, implies that (some) functional relationships among proteins, which could be expressed by associative relationships (see section 2.2.5) ore more advanced ontology features, cannot be expressed within the taxonomy. On the other hand, the restriction results in a much easier maintainability compared to the maintenance of a more expressive ontology. The right balance between simplicity and expressivity is subject to various discussions [SKK04].

4.4 Review with Users and Domain Experts; Refine Taxonomy

Before implementing the taxonomy, a review should be conducted to evaluate the taxonomy developed so far. Among evaluating formal criteria as addressed in section 2.1.2, the actual result should be tested against the requirements stated in the first development steps. Furthermore, issues to be considered when evaluating a taxonomy are:

- Usability, e.g. by having several users index the same content object, having several users define searches for this content object, and comparing the results
- The structure of the taxonomy by different (groups of) users
- Completeness
- Appropriateness of granularity, i.e. whether terms are missing or superfluous

Based on the evaluation, the taxonomy is refined according to the steps described above.

4.5 Implement and Test

Implementation and testing of the taxonomy with the applications for which it has been developed, preferably in one of the languages described in section 5, will not be treated within this document. Implementation aspects like display and functionality of taxonomies within electronic systems are comprehensively treated in [ANS05] and [BS05].

4.6 Manage and Maintain

Due to changes in the underlying domain, a taxonomy, unless it is deprecated, is never “finished”, but has to be continuously developed and maintained, which mainly refers to updating the terms and relationships. Updating terms comprises addition, modification, and deletion of terms.

Addition of a term is necessary when indexing or searching reveals terms not yet in the taxonomy. New terms for indexing may be necessary when new content objects have to be described by the taxonomy. The techniques described above, e.g. (semi-)automatic detection of appropriate term candidates may be applied. On the other hand, additional term candidates may be identified by evaluating search logs recorded from user queries [ANS05]. Additional relationships have to be defined according to the steps described above.

Indexers and searchers may propose modifications of terms according to defined policies and procedures. The original terms should be kept with the taxonomy as non-preferred synonyms (see section 2.2.1) within the taxonomy, as they may have been used for indexing.

According to [ANS05], overused terms and terms not frequently used for indexing should be considered for modification or deletion *"as both kinds of term are generally ineffective in retrieval"*. Deleted (as well as modified terms) may be kept in a taxonomy for retrieval purposes, being marked as "deleted". Relationships of deleted terms have to be considered appropriately. In case of terms with related terms, these have to be deleted as well, or they have to be rearranged. Consideration also has to be given to re-indexing current content objects.

Additionally, addition, modification, and deletion of terms have to be supervised by a control board that has to decide about proposals concerning these operations. As it is not practical to promote each of these operations immediately, they have to be subject to a version control and release management.

A crucial issue is the dissemination of modifications of a taxonomy. According to [BS05], the simplest way of disseminating the changes is to distribute a whole new taxonomy, but there may be target systems that link taxonomy data to other data. In this case, as a minimum, reports or files containing the changes should be provided.

4.7 Organisation Structure

According to [AW80], for the development of taxonomies three kinds of organisational structure are possible[5]:

- A centralised structure. All decisions are taken by a central agency, other agencies contribute terms and suggestions.
- A decentralised structure. Each of the cooperating agencies assumes *"responsibility for selecting and interrelating the terms which fall within its own language and/or subject areas."*
- A semi-centralised structure. *"The work is controlled by a central editorial committee consisting of delegates from the various cooperating agencies."*

According to [AW80], a centralised structure allows fast decisions, but may not take the views of the other cooperating agencies sufficiently into account. A decentralised structure is regarded as least effective. A semi-centralised structure is regarded as the most effective, which applies to a federated development in the monolingual case as well.

4.8 Multilingual Taxonomies

All of the construction steps of the development process as well as the principles and best practices introduced in the preceding sections apply to multilingual taxonomies as well. Like monolingual taxonomies, multilingual taxonomies can be constructed from scratch or by reusing (one or more) existing taxonomies. The only additional construction approach, a special case of reuse, is the translation of an existing taxonomy into (one or more) other languages.

[5] In fact, [AW80] states these kinds of organisation structure for multilingual thesauri, but they apply for each kind of taxonomy as well.

For each of these approaches, [AW80] proposes to allow feedback, i.e. to change the form or structure of a term in one language to achieve an easier or more useful solution in another language. For example, a compound term as the German "Lehrerbildungsgesetz", for which there is no equivalent English or French term except from a complicated paraphrase as "Law of education of teachers", may be factored into the terms "Lehrer", "Bildung" and "Gesetz", which will be much more useful for non-German users. The compound term may be included as loan term in the non-German language versions of the taxonomy.

4.9 Extension of a Taxonomy

In some areas of public administration, a superordinate taxonomy is provided that must not be changed except from extensions limited to specific parts of the taxonomy. For example, NACE is the official classification for economic activities within the EU. The only allowed extension of NACE is an additional level for branches specific to a Member State.

Extending such a taxonomy corresponds to the reuse of an existing taxonomy apart from the following differences:

- The development activities including maintenance are restricted to the allowed extensions.
- Concerning maintenance of the taxonomy, changes of the superordinate taxonomy may impose changes on the extensions.

Thus, the construction and maintenance process of an extension of a taxonomy is not different from the regular case.

4.10 General Recommendations

In general, it is highly recommended to utilise the SEMIC.EU platform and the SEMIC.EU services during the whole life cycle of a taxonomy, i.e. in each step of the construction process as well as in the maintenance phase, and for each case of taxonomy development, e.g. reuse, extension, etc.

In detail, it is recommended to:

- Register the project right from the start in order to achieve visibility
- Look for existing taxonomies in order to reuse them
- Start a forum thread and launch a call for comments to inform the interested community and to find collaboration partners
- Continuously publish new versions of the taxonomy and take part in SEMIC.EU's maturity process
- Continuously keep in touch with the community by communicating new requirements, consideration, design decisions, etc.
- Possibly achieve to pass SEMIC.EU's conformance process

Following these recommendations will increase the potential of saving costs, in particular, when reusing existing solutions, meeting user expectations and increasing the quality of the solution by utilising SEMIC.EU's maturity and conformance process.

This section introduced principles, techniques, and good practices concerning the life cycle of a taxonomy, covering the whole process of development and maintenance of taxonomies in detail, including multilinguality and the extension of a taxonomy. In the following, the integration of heterogeneous taxonomies will be presented. The next section will deal with the syntactic integration, i.e. the integration of taxonomies written in different languages.

5. TAXONOMY LANGUAGES

The overall aim of SEMIC.EU is to support interoperability between and among its European members, which means that it enables effective and efficient electronic data exchange of all artefacts necessary for them to reach their business goals. In the scope of these guidelines, the focus is exclusively on artefacts of the type “taxonomy”.

Electronic data exchange implies that taxonomies are represented in machine-readable form, in contrast to taxonomies drawn on whiteboards as a basis for discussions between engineers, for instance. When machine readability is one of the requirements, one needs to employ a formal language (sometimes called ‘format’) to represent those artefacts, namely a formal language for taxonomies.

The present chapter deals with the exchange and integration of taxonomies on the syntactic level. For this purpose, the next section discusses basic questions that need to be answered when choosing a language. Background technologies and advanced techniques for representing taxonomies are presented in section 5.2 and 5.3, respectively. Finally, relevant standards and the architecture for syntactic integration based on these technologies are described in section 5.4.

5.1 Basic Issues

The primary language-related issue is whether a new language should be developed or the use of a standardised language is to be preferred. Developing a new language for taxonomies has several observable disadvantages, the first being the concomitant requirement to develop language-dependent tools. Examples of such tools are parsers, transformation components, and visualisation components. Transformation components play a particularly crucial role in the exchange and adaptation of taxonomies.

For instance, if a spreadsheet format is chosen, transformers to other formats have to be developed because well-defined standard transformation components like XSL cannot be used. It is also not possible to rely on software frameworks that offer this type of functionality, and so parsers have to be implemented instead.

An example of a standard software solution is the case of Simple Knowledge Organisation System SKOS[6] [MB09], where XML-based frameworks are able to parse the XML representation of SKOS models. Another advantage of applying XML frameworks is that they support the presentation of taxonomies in the form of strict trees within visualisation components.

Therefore it is not advisable to rely on highly customised taxonomy languages when exchanging taxonomies. Especially in the context of interoperability projects under the roof of SEMIC.EU, it is a big advantage with respect to efficient work processes that the barrier to understanding taxonomies and other artefacts developed by European partners is as low as possible. The agreement on commonly used notations and languages will support such goals.

Thus, generally speaking, the use of standards is preferred. However, recognising that circumstances sometimes require pan-European collaborating partners to use their own languages, this document and [FI09] will provide information about these standards in a way that enables these partners to identify the standard for which syntactic mappings (see section 5.4) have to be defined.

[6] In this context, we focus on pure SKOS and do not consider its ability to integrate Description Logic [BCM+03] constructs from OWL.

5.2 Background Technologies

The two most relevant fields with respect to taxonomies are the knowledge-representation field of the Artificial Intelligence community [BL04] [SS08], which dominates the area of semantic technologies in general and the semantic-web branch of the W3C [BHL01], and the field of content classification and search of the community of digital libraries and information science [SS08].

Core technologies and languages concerned with content classification and search are given by thesauri and classical controlled vocabularies [ANS05]. Core technologies and languages stemming from the area of knowledge representation are divided into two groups: one supporting logical rules [CL07] and another supporting terminological expressions and axioms [BCM+03].

The discussion above shows that we have to assume the use of different approaches for taxonomy languages. Of course, this has a huge impact on the interoperability issues that SEMIC.EU has to deal with. This means that before we can consider semantic interoperability, syntactic interoperability issues caused by the different approaches described above have to be clarified. Therefore, this chapter will consider syntactic interoperability, whereas semantic interoperability will be discussed in section 5.

The different core technologies mentioned above offer a wide range of functionality and expressivity. This functionality is of obvious benefit in regard to semantically rich vocabularies. Specifically, in the context of the (semantic) web, it is useful to have this kind of support because data structures of different sites have to be exchanged.

For instance, one of the main features of highly expressive languages like OWL is that they provide constructs that enable vocabulary designers to define information models based on schema modelling primitives (for details see [FI08]). These primitives are not relevant, though, when only taxonomic structures have to be dealt with.

Concept hierarchies are formed via the generalisation/specialisation relationship. But not only generalisation/specialisation has to be taken into account concerning the interrelationships between concepts within taxonomies. For instance, the thesauri standard [ANS05] knows the associative RT (‘related term’) relationship that declares two concepts as semantically similar.

Furthermore, it is also a big advantage to enrich the taxonomic structure with associative relationships that are identified by a name. In the next section, an example is given for the usefulness of this language element.

5.3 Extensional and Intensional Definitions of Taxonomies

Due to the fact that many partners with varying technological backgrounds are involved in a SEMIC.EU project, many kinds of languages or formats for taxonomies with different levels of expressivity are present. One of the most essential aspects of such languages is the distinction between extensionally or intensionally defined taxonomies.

This distinction is critical for the syntactic integration of taxonomies because it has to be reflected in the selection of standards as well as in the integration architecture. Specifically, the standards chosen have to cover both techniques and must be connected by the integration architecture. This point is addressed in section 5.4, while the two techniques are described in the following.

Before applying these techniques to taxonomies, the example of simple sets is offered to demonstrate the difference between an intensional and extensional definition. Let us assume that we want to define the numbers one to five. Extensionally, these numbers are given by the set *{1,…,5}*. One possible intensional definition could be the statement ‘all natural numbers that are greater than zero and less than six’.

In the context of controlled vocabularies, ‘extensionally’ means that one explicitly lists all concepts and all relationships between these concepts. It is important to recognise that classical taxonomies are extensionally designed. Intensional definitions describe the properties of things and relations in order to enable the ‘reader’ of the definition, which can also be a machine, to derive the facts that are meant.

For instance, consider the following intensional definition:

(a) Every *person* that legally drives a *car* is an *adult driver*.

Here, we do not enlist all adults in the world explicitly. Instead, we give an intensional definition for adults. Assume additionally the following definition:

(b) Every *trucker* legally drives a *truck*.

Together with the explicitly defined specialisation (c) *truck* ⊂ *car*, it follows that (d) *trucker* ⊂ *adultDriver* ⊂ *person*. In this sense, the intensionally defined taxonomy consisting of (a), (b), (c) stands for the extensionally defined taxonomy (d).

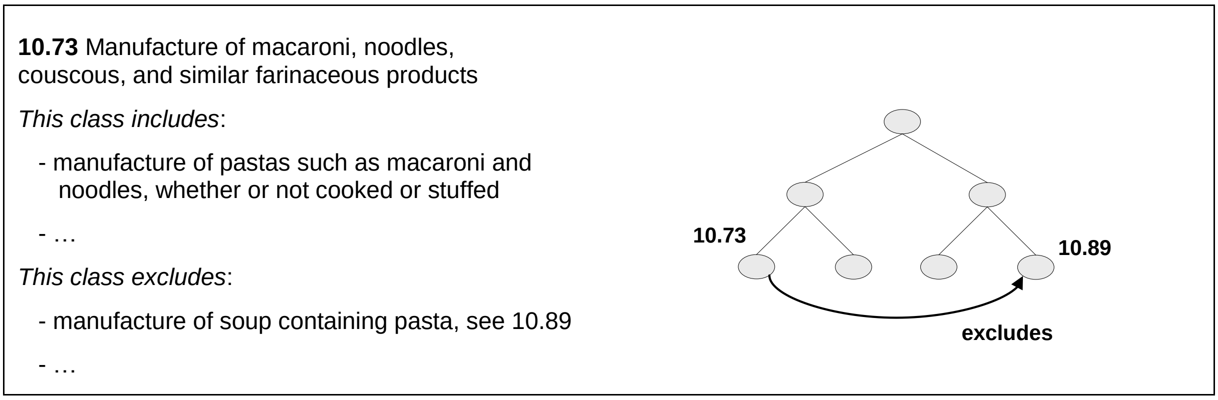
In the context of SEMIC.EU, the example above is part of a realistic scenario for exchanging electronic artefacts, which in our case are artefacts of type taxonomy where one collaborating partner offers a taxonomy that is partially based on intensional definitions in taxonomy language *A* that offers such expressivity. Partners who want to reuse this artefact in their systems and who have the constraint that their systems only accept taxonomies in thesaurus format *B* have to be provided with tools to transform A to B syntactically.

Clearly, this is not a pure one-to-one transformation, which will be described later in this section. Note that no heterogeneity of taxonomies is involved here; handling heterogeneous taxonomies will be discussed in section 5.

The above-mentioned language example shows the usefulness of terminological expressions in the form of (a) and (b). Besides this kind of advanced definitions, some approaches admit the possibility of defining taxonomies via simple rules. For instance, the simplified rule

- car *x* drives faster than 250 km/h and has only 2 seats ⇒ *x* is a sports car

together with the fact that cars of class Porsche Carrera fulfil the rule premise, the specialisation *PorscheCarrera* ⊂ *SportsCar* holds.



10.73 Manufacture of macaroni, noodles, couscous, and similar farinaceous products

This class includes:

- manufacture of pastas such as macaroni and noodles, whether or not cooked or stuffed
- …

This class excludes:

- manufacture of soup containing pasta, see 10.89
- …

Figure 4: Exclusion in NACE

Whereas thesauri languages (or standards) support the extensional definition of taxonomies, the advanced technologies and languages mentioned above allow for intensional definitions. To reiterate, these guidelines focus exclusively on taxonomies. Therefore, only the part of these advanced technologies that targets the definition of taxonomic structures will be considered.

A practical example of how to enrich taxonomic structures by means of advanced approaches, namely rules and terminological expressions, is the definition of constraints for the NACE taxonomy [NAC08]. In the definition of NACE, which is of course a taxonomy documentation rather than a machine representable format, every class is not only specified by a description of the elements the class has to include, but also of those elements the class has to exclude.

Exclusion cannot be fully expressed by a classical taxonomy language. There are two expressivity levels for specifying this constraint: (1) The exclusion can simply be stated as an association with the name ‘exclude’, but this statement would only be informative, and (2) this statement can be modelled by a terminological expression or rule.

The first expressivity level is exemplified on the right side of Figure 4. It shows a fragment of the NACE taxonomy where a class concerning food manufacturing is declared (class 10.73) by describing activities that belong to this class and by excluding activities that belong to a different class (10.89). On the left side, the taxonomic structure is stated with a named association (‘excludes’).

Whereas the first level of expressivity can be specified with a classical taxonomic structure (supported by extended thesauri), the second level is only expressable with terminological expressions or rules. The corresponding terminological expression and the same fact with rules for the above example are

- *class(10.73)* $\subset$ *not class(10.89),*
- *x in class(10.73) and x in class(10.89)* $\Rightarrow$ *failure,*

respectively. These expressions can be interpreted by the following: Everything that is within *class(10.73)* cannot be within *class(10.89)* and vice versa.

Indeed, the thesauri standard, extended by named associative relationships, is able to represent the first expressivity level. But the second level, which allows for automatic validation of the taxonomy concerning the exclusion constraint, is only supported by languages that provide constructs for terminological expressions or rules.

5.4 Syntactic Integration

Syntactic integration has to be oriented toward the dimensions described in the previous sections, which means the background technologies as well as the distinction between extensional and intensional techniques have to be reflected by the standards and the integration architecture chosen. The next section discusses the standards that support all language aspects introduced so far. Finally, the architecture for syntactic integration will be presented in section 5.4.2.

5.4.1 Standards as the Key Languages

The above discussion of the key aspects of languages for representing taxonomies has shown that three kinds of taxonomy languages (distinguished by their respective core technologies), namely

- classical thesauri-oriented languages with explicit statements
- languages with terminological expressions
- languages with rules over concepts

play an essential role in the context of SEMIC.EU, where pan-European collaboration is based on a large diversity of approaches for representing taxonomies. As described, the first kind allows for extensional definitions, whereas the last one deals with intensional ones.

Many systems meeting such requirements have been designed and developed in the past few decades that allow for taxonomy definitions in a classical sense or by means of rules and terminological expressions [ANS05] [IFLA09] [CL07] [BCM+03]. Because of the nature of the SEMIC.EU project, where electronic data exchange via the internet is required, the document has to focus on standards where web-oriented approaches are of special interest.

In order to achieve the goal of syntactic integration based on standards, it is necessary to have one standard for each core technology introduced so far. The standard for explicitly defined taxonomies (extensional) is given by SKOS, whose clear advantages are described in [FI09]. It offers all relevant features based on the thesauri standard [ANS05].

Without doubt, the standard for the core technology of terminological expressions and axioms is Ontology Web Language (OWL) [BHH+04]. But because we do not focus on (full-fledged) ontologies in this document, the full range of OWL functionality is not considered here; only features that take care of taxonomic structures are of interest. OWL can be seen as highly elaborated web customisation of Description Logics [BCM+03]. Furthermore, it reflects the state of the art within this community.

There are at least three candidates, namely Rule Interchange Format (RIF), which is strongly influenced by F-Logic [KLW95], Web Service Modeling Language (WSML), and Common Logic, for the third core rule-based technology [RIF09] [WSML08] [CL07]. Common Logic is very strongly related to the Artificial Intelligence community and is of high expressivity compared to full predicate logic. Because of the fact that full predicate logic is far beyond the constructs for taxonomic structures, the first two languages are more suitable candidates.

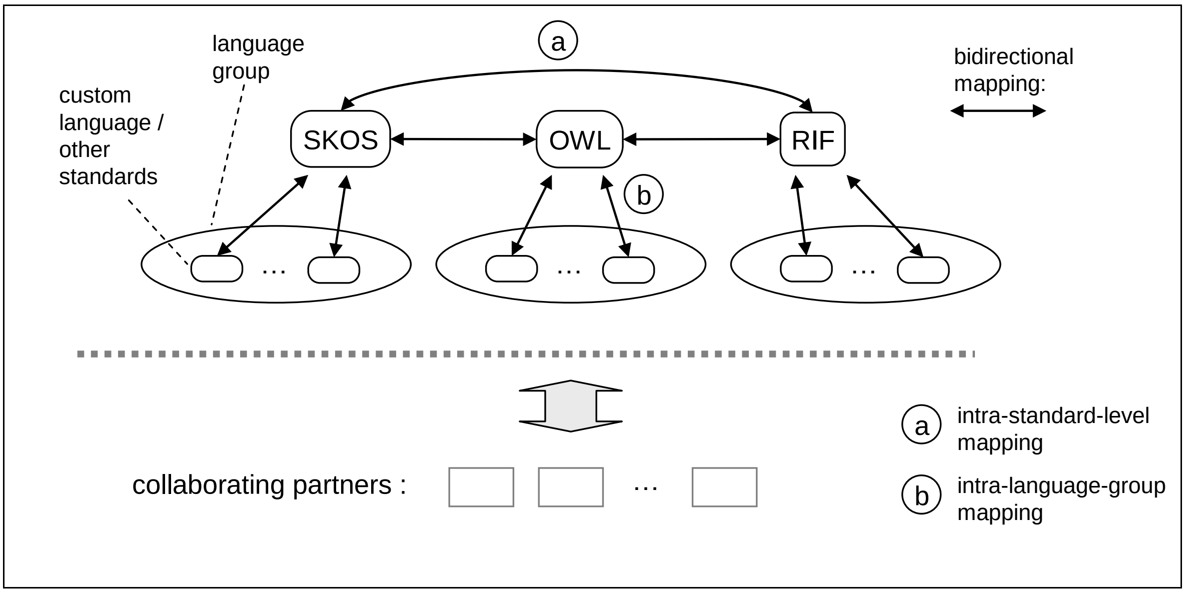

language group
custom language / other standards
a
SKOS
OWL
RIF
b
bidirectional mapping:
…
…
…
collaborating partners :
…
a intra-standard-level mapping
b intra-language-group mapping

Figure 5: Syntactic Integration Architecture

WSML was initially planned for tasks in conjunction with web services, but it is also suitable with respect to the modelling of taxonomies. Workings concerning WSML are reflected by the RIF work, which can be seen as an amalgamation of many praxis-relevant systems. Therefore, RIF can be chosen for representing taxonomies that make use of a rule part to intensionally define concepts.[7]

As mentioned, many customised taxonomy representations are used in national and European projects ranging from Excel tables to SKOS instances. Of course, a bidirectional syntactic mapping between these would be simpler than a semantics-preserving mapping because only syntactic structures are transformed. But the same problem of combinatorial explosion occurs when many different formats

[7] It has to be noted that the SEMIC.EU portal contains an asset with a WSML artefact. RIF is currently work in progress but is soon expected to become a standard.

are used. An integration architecture that considers such problems will be discussed in the next section.

5.4.2 Integration Architecture

The use of standards for syntactic integration directly follows the pivot paradigm discussed in [FI08] to avoid a huge amount of bidirectional mapping.

Figure 5 shows the general architecture required for transforming between different languages based on the three standardised languages that serve as syntactic pivot elements. On the first level, the three standards are shown together with definitions for their bidirectional mappings.

Principal use case	Mapping application
The partners use the same taxonomy language.	No mapping has to be applied. If the project is expected to involve other languages, mapping to the corresponding standard should be provided.
The partners use taxonomy languages within the same group.	Only mappings within one group have to be applied. This will be done via the standard related to the group.
The partners use taxonomy languages from different groups.	Mappings to the standard of the group followed by mappings between standards and again mappings to the group have to be applied.

Table 3: Principal Use Cases

On the second level, the figure shows three groups of languages. For instance, the left group contains all custom languages that are representable one-to-one by SKOS. The last level symbolises the project partners who use a custom language or a standard. These partners have to electronically exchange taxonomies as instances of the corresponding language or format.

Depending on the level, a distinction has to be made between intra-standard mappings and intra-language-group mappings. Whereas the first transform between the standards, intra-language-group mappings define how to transform from one custom language to another within the same group. Transformations between custom languages within one group are performed by using the mapping into the respective standard followed by the mapping to the target custom language.

In this sense, we can speak of two-level-pivoting because within one language group, the pivot element is the respective standard and between two different groups, the standard triumvirate itself can be seen as the pivot element.

From the architecture in Figure 5, it follows that different use scenarios for SEMIC.EU projects involve different sequences of mappings between taxonomies. Table 3 summarises different mapping situations by means of three principal use cases. Concrete use cases will be a combination of these three applications.

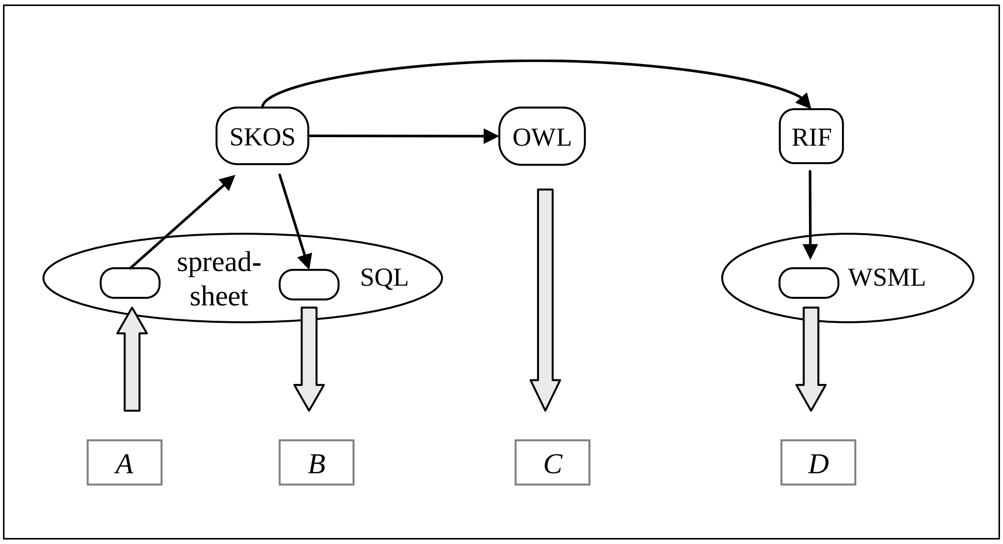


Figure 6: Example Transformation

Figure 6 shows a concrete scenario conforming to the architecture in Figure 5. The project partner *A* uses a custom taxonomy language in tabular form with an explicitly defined specialisation hierarchy (spreadsheet file). The other project partners, who want to use the taxonomy of *A*, have systems that base on a customised SQL database for the specialisation hierarchy, OWL, and WSML, respectively.

According to the above example and Table 3, the following mapping applications are performed. First, the spreadsheet file is mapped to the standard representing the first group, namely SKOS. For the partner using the SQL database, the result is mapped to respective database tables (intra-language-group mapping, see Figure 5). The remaining partners use languages belonging to other groups.

Therefore an intra-standard mapping from SKOS to OWL and RIF is applied. Finally, the RIF instance is mapped to WSML.

Principally, taxonomies in a custom language with an explicitly defined specialisation hierarchy (first group) can be mapped to SKOS directly. Mapping from languages, following the intensional paradigm, to SKOS via OWL and RIF cannot be performed directly. In such cases, reasoning and normalisation techniques have to be performed [VS09].

Whereas the present section introduced the syntactic integration of taxonomies, i.e. the mapping of a taxonomy represented in one language to a taxonomy represented in another, the next section will introduce integration of taxonomies that are semantically heterogeneous. Related techniques for detecting heterogeneity and their subsequent use for merging, translation, and mediation of taxonomies are introduced.

6. MATCHING, MAPPING, AND INTEGRATION OF TAXONOMIES

One of the important scenarios of SEMIC.EU is that several European Member States want to exchange documents concerning their citizens or that those citizens themselves want to use their locally created documents in other European countries. For instance, a citizen of one Member State wants to work in another Member State based on personal skill profiles. These skill profiles consist of terms stemming from a skill taxonomy used in his country. Without a valid exchange of terms, it will not be possible to meet the cross-border needs of European citizens since this type of exchange assumes that terms used to classify data are well-understood in every European country.

Therefore, taxonomies as essential artefacts for clarifying semantics of terms play a central role within SEMIC.EU projects. In order to ensure interoperability between local taxonomies, it is necessary to apply integration techniques such as taxonomy matching and mapping.

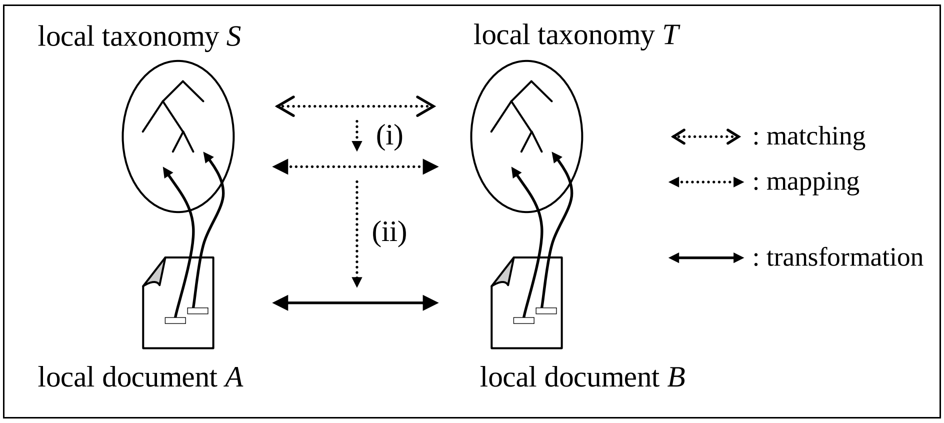


Figure 7: Exchanging of Taxonomies

Figure 7 depicts this kind of scenario on an abstract level. The local document *A* of an entity (i.e. an organisation or citizen) located in one Member State has to be transformed to a semantically equivalent document *B* for the usage in another Member State. All terms stated in *A* stem from the local taxonomy *S*. But according to the rules of the other Member State local taxonomy *T* is used for classifying the corresponding domain. Therefore, the terms have to be correctly mapped from *S* to *T* in order to guarantee the comprehension of the document.

For the purpose of creating these semantic-preserving mappings (see (i) in Figure 7), various matching techniques between the local taxonomies are applied. Mapping definitions in turn control the correct transformation of document content (ii). These mapping definitions, which are possibly arranged according to pivot structures, are the corner stone for successful semantic integration of taxonomies locally defined in different Member States.

In contrast to syntactic integration discussed in section 5, integration in the context of the present chapter is complicated by semantic heterogeneity. Generally, semantic heterogeneity occurs when two or more designers use different structures and naming conventions to model the same domain. Due to the fact that taxonomic structures are the most essential part of ontologies [NM01a], all approaches to the integration of heterogeneous ontologies consider the integration of taxonomic structures in particular.

Therefore, these guidelines will consider concepts and solutions from this community as far as possible [Noy04] [ES07]. Of course, many problems treated by the ontology community are not relevant with respect to taxonomies. In particular, integration problems that stem from heterogeneous schemata and their solutions based on extensive experience in the schema-integration community are not considered here in conjunction with pure taxonomies [BLN86] [SL90].

The present chapter is structured as follows. The first section gives a conceptional introduction to the interrelations between domains, heterogeneity, and upgrading. Taxonomies are used to model domains and, depending on the shared domain parts, heterogeneity can occur. Section 6.2 discusses kinds of heterogeneity that can be observed when two taxonomies partly model the same domain.

Section 6.3 addresses taxonomy matching approaches to be used for the detection of semantic heterogeneity. Semantic equivalence as the essential relatedness of concepts is discussed and different kinds of semantic equivalence are presented. To represent different kinds of semantic equivalences, correspondences and alignments are introduced. Furthermore, the general structure of the matching calculation for determining these alignments is presented.

Section 6.4 deals with the semantic integration of taxonomies. Based on the results of the matching calculation, i.e. alignments, various kinds of taxonomy mappings are introduced. Mappings are used to merge taxonomies as well as to translate concepts from source to target taxonomies. Furthermore, a mediation scenario is presented that makes use of constrained mappings in order to guarantee semantics preservation. Finally, aspects of work processes are discussed, and their interrelation in the efficient integration of heterogeneous taxonomies is described.

6.1 Domains, Heterogeneity, and Upgrading

As described in [FI09], vocabularies in the sense of dictionaries usually refer to domain-independent term lists in which grammatical information is also sometimes available. Taxonomies as a special type of controlled vocabularies are used to describe terms from a specific domain.

How specific the taxonomy has to be depends on the field in which it is to be applied. With respect to domains, a distinction exists between ‘upper taxonomies’ (or ontologies) and domain taxonomies [NP01]. Whereas upper ontologies take care of general terms like ‘event’, ‘action’, etc., domain ontologies deal with concrete domains like health, financial concerns, etc.

In this sense, upper taxonomies model the universal domain. Within the practical work of electronic data exchange in the context of SEMIC.EU, domain taxonomies and their relationship to heterogeneity are of higher interest. Although issues also occur at the level of upper-taxonomy heterogeneity, domain taxonomies have a higher susceptibility to heterogeneity because a much higher degree of agreement exists among designers concerning the modelling of the universal domain than modelling a specific domain.

One of the main unknowns is in regard to the point(s) at which semantic heterogeneity of domain taxonomies has to be expected. The answer is directly connected to the disjointness of the taxonomy parts with respect to the domain. Whenever (at least) two taxonomies model the same domain, there is a high degree of semantic heterogeneity. More precisely, the grade of heterogeneity expected depends on the grade of overlapping of the taxonomies with respect to a modelled (sub-)domain.

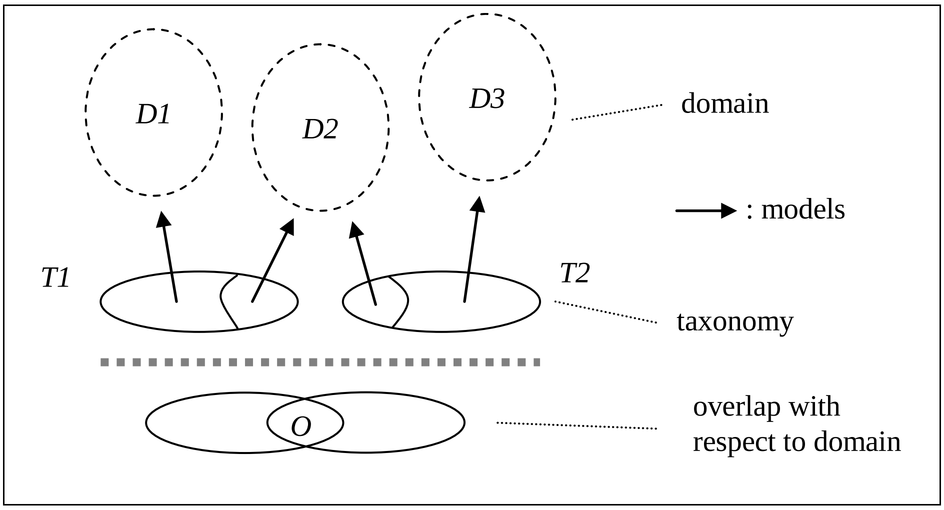


Figure 8: Taxonomies Sharing the Same Domain

Figure 8 shows this coherence schematically. There are three domains *D1* to *D3* and two taxonomies *T1* and *T2*. *T1* and *T2* mainly model the domain *D1* and *D3*, respectively. Domain *D2* is modelled by both taxonomies to a certain extent. Here, an overlap is given with respect to the modelled domain *D2,* and therefore semantic heterogeneity can be expected.

For instance, a home improvement store and a furniture store are modelled by means of two taxonomies. The taxonomy for the former covers all concepts relevant to home improvement like drill machines, nails, and paint. Furthermore, the taxonomy models products from the bathroom department, such as toilet seats, mirrors, etc. The furniture store also models bathroom furniture as a part of its furniture selection, which contains all the furniture used in a home. The common domain modelled by these two taxonomies is bathroom furniture.

Of course, semantic heterogeneity has a strong impact on the integration of taxonomies. If several taxonomies are disjoint under the domain view, no semantic heterogeneity exists, and the main task in integrating them is to interrelate their concepts (RT; related term), assuming that thesaurus expressivity has been chosen. In the case of a taxonomy language with named associations, relationships can carry more semantics.

Another kind of integration, which does not take semantic heterogeneity into consideration, is called ‘upgrading’ [SS08].[8] In this context, upgrading is accompanied by an increased level of expressivity. The upgrading process is geared to the levels of construct sets that are available on the specific expressivity level. The first step is the introduction of generic associations represented by the RT construct. The second step refines these associations by means of named associations.

[8] Upgrading taxonomies is linked to the topic of what are termed ‘crosswalks’ between knowledge systems. This concept arises in the context of knowledge integration from the information science perspective.

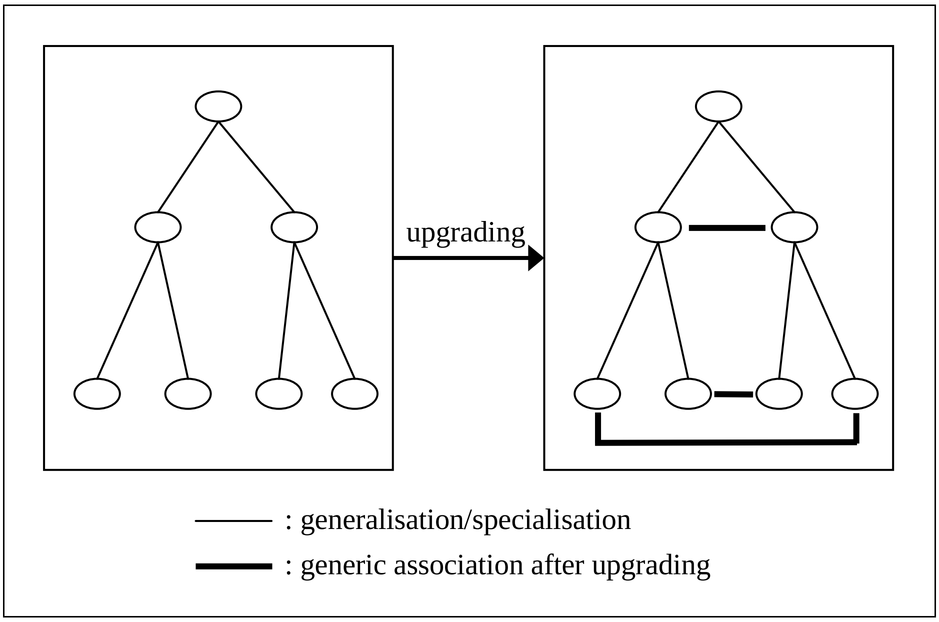

upgrading
: generalisation/specialisation
: generic association after upgrading

Figure 9: Upgrading a Taxonomy

Figure 9, adopted from [SS08], shows the first step schematically; here siblings are connected by generic associations. The second step exhibits a semantic enrichment and can be demonstrated by the following example based on a thesaurus fragment:

- **research project**
  related term: research institute
  related term: funding agencies

The upgraded version could be given by

- **research project**
  hostedBy: research institute
  supportedBy: funding agencies

where the RT relation is refined by the properties *hostedBy* and *supportedBy*.

In the context of SEMIC.EU, upgrading taxonomies can play an essential role with respect to the tools and systems used by collaborating partners. Above, the only action performed is the extension to thesauri, including named associations. Although this extension is a small step in the context of the structure of the vocabulary, it has many advantages from the tools and systems perspective.

Namely, tools and systems whose aim is to support indexing, search, and visualisation benefit from associative relationships between concepts [SHS00]. The second level of upgrading is the enrichment of taxonomies with property definitions. Property definitions, and more general schema constructs, make an integrated view of information models and terminologies possible. This is particularly useful for describing semantics of schema elements by referring to items in the terminology.

The third level of upgrading concerns techniques stemming from full-fledged ontologies. That is, rules or concept expressions serve to express constraints for the taxonomy that permit automatic validation of the taxonomy. The three levels of upgrading are summarised in Table 4.

Principal use case	**Upgrading level**
The partners want to use tools for indexing, search, and visualisation.	Taxonomies should be enriched by ‘related term’ associations or by named associations.
The partners want to have an integrated view of schemata and vocabularies.	Taxonomies should be enriched by property definitions.
The partners want to use tools and systems for consistency checks or validation.	Taxonomies should be enriched by rules or terminological/concept expressions that specify how concepts are restricted in relation to other concepts.

Table 4: Use Cases for Upgrading

The previous integration case assumed that the taxonomies are disjoint under the domain view. If taxonomies overlap under the domain view, heterogeneity has to be dealt with. Various kinds of heterogeneity are discussed in the next section.

6.2 Kinds of Heterogeneity

As described in the previous section, heterogeneity stems from the overlap of two or more different taxonomies vis-à-vis the (sub-)domain they model (see overlap *O* in Figure 8). The present section describes the kinds of heterogeneity that can occur in this situation. These serve as a conceptual base for the development of techniques to detect heterogeneity that are discussed in section 6.3.

In the area of ontology matching, one of the most exhaustive descriptions that deal with heterogeneous parts and semantic correspondences between ontologies is given by [ES07]. Although it focuses on ontologies as the most expressive language, the results can be reused because taxonomic structures are the essential building block of ontologies.

According to research in ontology and taxonomy matching, three kinds of heterogeneity generally prevail within the community:

- Syntactic heterogeneity
- Terminological heterogeneity
- Conceptual heterogeneity.

Syntactic heterogeneity deals with the situation where two taxonomies are represented in different languages. Syntactic heterogeneity is considered in section 5.

Terminological heterogeneity can be observed when terms that are used for lexical representations of a concept differ. More precisely, the same concept can have different lexical representations in two taxonomies. For instance, terminological heterogeneity occurs when two taxonomies use synonymous terms for the same concept. Detecting terminological heterogeneity can be performed using techniques on the node or element level introduced in section 6.3.3.

Conceptual heterogeneity is the most complex form of heterogeneity [ES07]. In this case, heterogeneity can stem from the fact that different taxonomies model the domain with a different level of concept granularity. For instance, one taxonomy models ‘professions’ on a fine-grained level by specifying ‘teacher’ with its different species ‘lecturer’, ‘professor’, and ‘docent’. The other taxonomy only considers ‘teachers’.

Structural granularity can be another source of conceptual heterogeneity. In this case, all bottom concepts, i.e. all concepts that are not specialised, are more or less the same in two taxonomies, the only difference being with respect to the specialisation structure. For instance, in the context of a supermarket, one taxonomy can structure the concepts ‘banana’, ‘apple’, and ‘beans’ with the super-concepts ‘fruit’ and ‘vegetable’, whereas another taxonomy does not incorporate the latter.

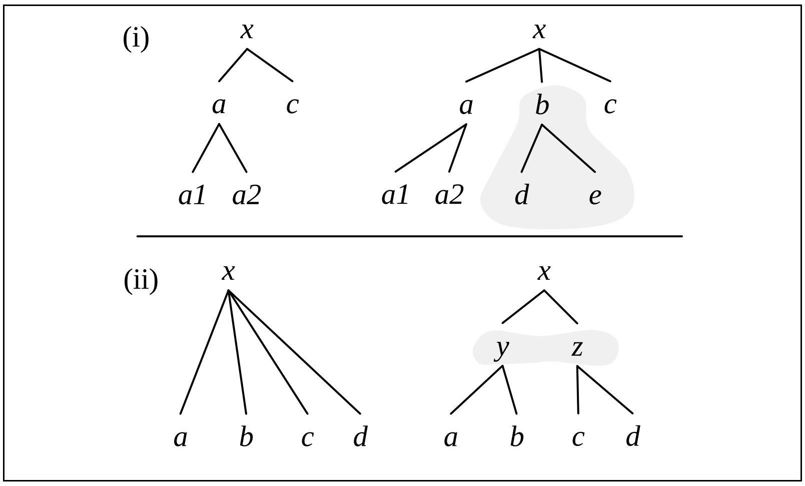


Figure 10: Concept Granularity and Structural Granularity

Concept granularity and structural granularity as two sources for conceptual heterogeneity are schematically exemplified in Figure 10. Part (i) shows concept granularity. On the left side, a taxonomy fragment of the first taxonomy is given. The right part presents the second taxonomy with a higher concept granularity due to the introduction of three new concepts *b*, *d*, and *e* (depicted by the grey region).

Part (ii) of the figure shows heterogeneity stemming from different structural granularities. The left side shows a relatively flat structure with four bottom concepts. In order to refine the structure, which often occurs when too many species occur on one level, the right side introduces a new specialisation level consisting of concepts *y* and *z*.

Note that the extension of the second taxonomy in part (i), namely the set of things (or objects) in the world modelled by it, is increased, assuming that the concepts are disjoint. Increasing the extension disjointness is necessary because otherwise, the concepts introduced might cover things that are already included (*d* and *e* could be already in *c* and *a2*).

Part (ii) does not increase the number of elements in the extension (assuming that *y* and *z* only aggregate the bottom concepts and do not introduce new objects). Normally, heterogeneity stems from a mixture of these two basic forms, which also manifest interdependence: one can observe that the transition from the left to the right fragment in part (i) accompanies the increase of structural granularity shown in part (ii).

6.3 Detecting Semantic Heterogeneity

One of the main questions is how to determine semantic heterogeneity. According to Figure 8, this has to be ascertained by determining the domain-related overlap between different taxonomies. But clearly, *O* in Figure 8 is not given explicitly.

The determination of *O* can only be made by relating the entities of the two taxonomies. It has to be noted that in our case, i.e. in the case of the restriction to taxonomies, the main entities are the concepts given by the nodes within the taxonomic hierarchy.[9]

6.3.1 Semantic Equivalence

Clearly, the border between the overlapping areas is not fixed, but rather 'floats' (see *O* in Figure 8). The reason for this floating border is that human beings can only make fuzzy judgements about equivalence or that automatic procedures based on comparisons of terms are involved (see section 6.3.3). The effect of these procedures concerning floating borders can be shown by means of the following example.

For the subsequent discussion, the following term pairs are given:

a) *(House, House)*

b) *(November, December)*

c) *(Car, Table)*

On the basis of lexical comparison, the relatedness between the terms of pairs a) and b) is the highest because of their being identical or the high degree of commonly used character groups, respectively. Pair c) is nearly unrelated because it contains no common characters.

The grade of relatedness between two terms described above can be seen as a confidence value describing the degree of equivalence between two concepts [SE05]. This confidence value is normally a number between zero and one. Other confidence values are possible, even simpler ones like discrete ranges (e.g. 'high', 'low', 'good', 'bad', or 'excellent'). The confidence value is calculated by the taxonomy-matching component according to the matching techniques used, which are described in section 6.3.3. With respect to the overlapping part *O* in Figure 8, floating borders directly correspond to these values.

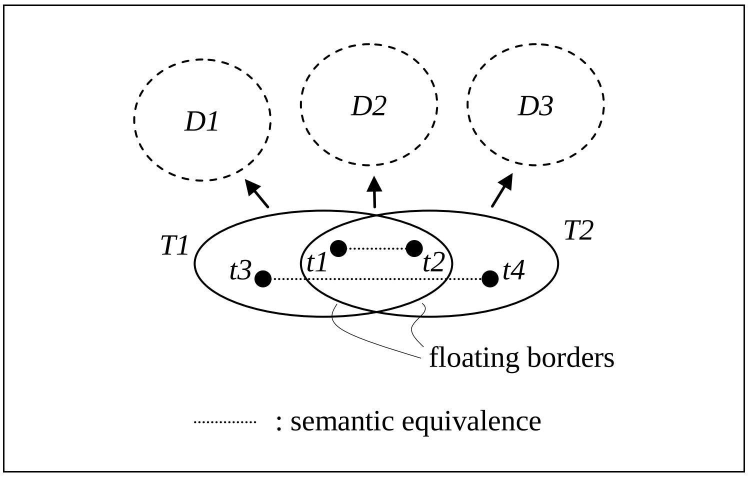


Figure 11: Semantic Equivalence

[9] Entities that also have to be considered in the case of schema matching are property nodes, data types, and data values.

Figure 11 shows the interrelation between the semantic equivalence of concepts and floating borders of the overlapping part vis-à-vis the modelled domain. As in Figure 8, two taxonomies are given that model two disjoint domains and one common domain.

Semantic equivalencies range from nearly non-equivalent term pairs like *(t3, t4)* from different domains to highly equivalent pairs like *(t1, t2)* from the overlap. The measure of equivalency, given by the confidence value, floats and corresponds directly to the borders of the overlap. The ideal situation is that a fixed confidence value *v* can be found (and with it fixed borders) such that all pairs with value *v* are in the overlap that models the domain *D2*.

So far, semantic equivalence between concepts (via their terms) has been exemplified by lexical relatedness. As a result, it appears possible to determine the grade of semantic equivalence given by a certain confidence value. Pure equivalence, however, is not the only consideration relevant for building pairs of concepts stemming from the taxonomies that have to be integrated.

The source of possible candidates for semantic relationships between concepts is given by the building blocks of controlled vocabularies themselves. In the case of taxonomic structures, the essential relationship candidate is generalisation/specialisation (here denoted by the symbol ⊂). Whereas above, a relationship was based on equivalence, e.g. resulting in a statement

- *(Paper, Article) is equivalent with a high confidence,*

now a relationship can have the form

- *concept c is a specialisation of concept d with a low confidence,*

where *c* occurs in one taxonomy and *d* occurs in another. Not only are building constructs of vocabularies candidates for a semantic relationship between two concepts of different taxonomies, but also elements of the semantics of taxonomies itself. As described in [FI08], taxonomies are interpreted by means of set theory. Besides the inclusion ⊂, the overlapping of sets is naturally a candidate.

Overlapping (or the inverse disjointness) is also reflected in some full-fledged languages: for example, OWL [BHH+04] contains the construct *owl:disjointWith*. In the following, the construct *m:overlap* is used to indicate the relationship of overlapping.

6.3.2 Correspondences and Alignments

Summarising the basic elements introduced so far, the concrete semantic relationship can be defined. This relationship is commonly called ‘correspondence’ [ES07]. Assuming two taxonomies *T1* and *T2*, the tuple

(t1, t2, r, v)

is called a correspondence where

- *t1* is contained in *T1,* and *t2* is contained in *T2,*
- *r* is a relation symbol for equivalence, specialisation, or overlap (=, ⊂, or *m:overlap,* respectively),
- *v* is a confidence value.

In the following, some examples of correspondences are given:

- *(Car, Automobile, =, high)*
- *(Engine, Car, m:overlap, middle)*
- *(House, Driver, m:overlap, low)*
- *(Biography, Essay, ⊂, 0.67)*

The first correspondence states that *Car* and *Automobile* are used equivalently with a high degree of confidence in the respective taxonomies. The next two correspondences state that *Engine*, *Car* and *House*, *Driver* overlap with middle and low confidence, respectively. The last correspondence predicates that *Biography* is a specialisation of *Essay* with a confidence value *0.67*.

Of course, the last statement is only concluded based on the two input taxonomies. If one were to start to design a new taxonomy, this relationship is not very realistic because biographies exist that are not in essay format. However, the two input taxonomies could have the following shape: All concrete biographies of the first taxonomy are located under 'essay' within the second taxonomy. Therefore, the specialisation *Biography* ⊂ *Essay* from the first to the second taxonomy can be concluded.

So far, the basic structures for semantic relatedness, namely correspondences between two concepts, have been discussed. In order to define the overall correspondence between two taxonomies, one has to consider all concept correspondences, commonly called 'alignment'. An alignment is defined as a set of correspondences between the concepts of the given taxonomies. Determining an alignment based on two taxonomies is called 'matching' and will be discussed in the next section.

6.3.3 The Matching Calculation

Commonly, taxonomy matching is defined as the calculation that determines an alignment.[10] Especially when large taxonomies (with thousands of elements) are used, it is very exhaustive to calculate the alignment between taxonomies manually. Therefore, tools have been developed in order to calculate these alignments (semi-)automatically (e.g. [NM01b] [GSY04]).

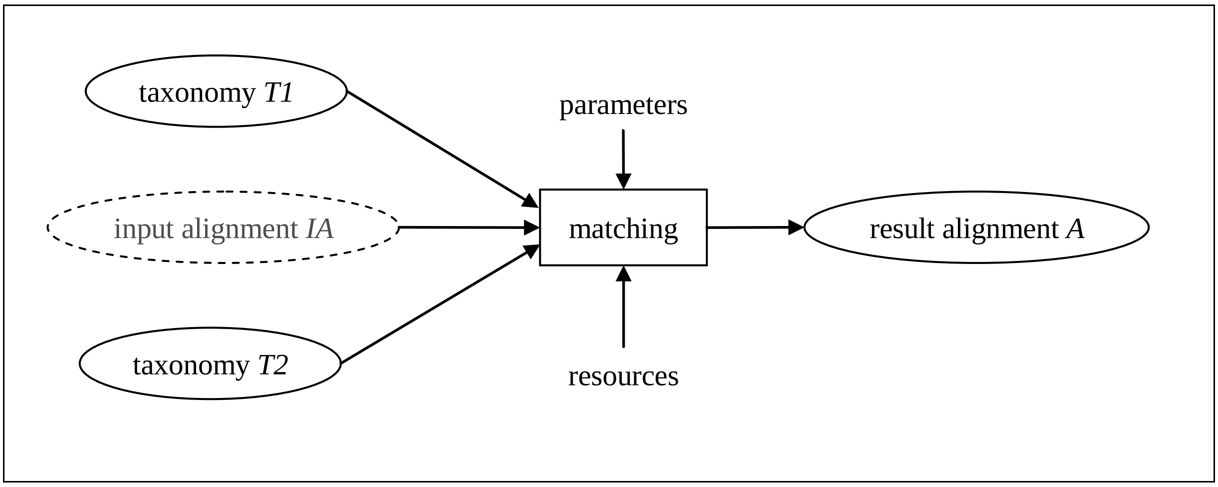


Figure 12: Matching Calculation

The general architecture of the matching component is shown in Figure 12 (cf. [BEF+04]). First, two taxonomies *T1* and *T2*, for which an alignment has to be found, are the main input of the matching. Next, parameters control the matching calculation. These parameters can consist of thresholds or directives that select preferred algorithms.

Thresholds for the confidence values can be set in order to determine the elements participating in an alignment. In other words, if the threshold is set to a value *v*, only correspondences with a confidence value higher than *v* will be taken into account (cf. Figure 11). Resources play an important role for

[10] In the following, we use the term 'matching calculation' rather than 'matching process' [ES07] in order to distinguish it from organisational processes or business processes.

elementary matching: For instance, dictionaries (with synonym rings) as resources support the matching on a linguistic level. The last input entity is also an alignment. Because a final alignment can be iteratively refined, the matching component supports already-defined alignments *IA* as input.

Matching techniques are presented in the following (cf. [SE05]). For a deeper discussion, we refer to [ES07]. It should be noted that most of these matching techniques work on extensional defined taxonomies (cf. section 5.3). If intensional defined taxonomies are considered, normalisation by means of reasoners has to be performed [VS09].

The top-level criterion for classifying matching techniques is given by the distinction between

- Element-oriented techniques
- Structure-oriented techniques.

Element-oriented techniques are applied on the nodes of taxonomies in order to detect terminological heterogeneity. More precisely, these techniques do not incorporate the taxonomic structure but focus on the lexical representation of concepts. Several commonly known classes of techniques are introduced below (for more detailed discussions, see [CRF03] [GSY04]).

String-based techniques assume that semantic similarity can be determined by calculating the similarity of character strings. Known algorithms are suffix and prefix comparisons, e.g. ‘phone’ and ‘telephone’ are assumed to be semantically near because the former is a suffix of the latter. Furthermore, ‘edit distance’ and ‘n-gram’ algorithms are applied.

Language-based techniques incorporate Natural Language Processing techniques. The most prominent technique is lemmatisation, which finds similarities by reducing terms to headword forms (e.g. houses → house). Techniques based on linguistic resources use other controlled vocabularies like dictionaries and thesauri for similarity determination (e.g. for synonym lookup). Finally, constraint-based techniques refer to comparisons of data types and multiplicities found in information models.

Structure-oriented techniques focus on the relationships between concepts. Several prominent classes of these techniques can be applied. The first one consists of graph-based techniques that make use of elementary graph-matching algorithms with special treatment of node children or graph nodes without outgoing edges [SWG02]. Matching techniques restricted to the generalisation/specialisation structure without considering associative relations take place in the case of pure taxonomies.

Case	Techniques	Top level criterion
(A)	String-based	Element-oriented
	Language-based	“
	Linguistic resources	“
	Taxonomic structure	Structure-oriented
	Repositories of structures	“
(B)	Graph-based	“
(C)	Constraint-based	Element-oriented
(D)	Model-based	Structure-oriented

Table 5: Suitability of Matching Techniques

More advanced techniques are model-based or incorporate external resources like structure repositories. The former techniques rely on formal, logic-based interpretations. In this case, taxonomies are encoded in logical statements like rules, which make it possible to calculate similarity by the application of reasoners. Also, Description Logic - as the basis of OWL - is applied in this case [BCM+03].

The application of the techniques introduced above depends on the expressivity used for specifying taxonomies. Although these guidelines focus on taxonomies, elements of more expressive languages are often used in practise for controlled vocabularies. Four cases (cf. Table 4) concerning the expressivity of the language can be distinguished from the viewpoint of matching techniques. Related constructs have been discussed in section 2 and [FI09]:

(A) Pure taxonomies

(B) Taxonomies enhanced by associative named relations between concepts (Thesauri)

(C) Taxonomies enhanced by schema elements (Ontologies)

(D) Taxonomies enhanced by logical rules (Full-fledged Ontologies)

As mentioned above, calculating alignments for very large taxonomies cannot be done without support by matching tools and systems. In SEMIC.EU, a realistic scenario could be that different partners use different expressivity levels ranging from (A) to (C). Table 5 is offered as support to partners in their decision regarding the technique(s) to be provided by the tool in order to calculate alignments of input taxonomies.

6.4 The Integration Phase

The output of the matching tool invocation is an alignment that should only be used for subsequent integration tasks after having been checked by a taxonomy developer and domain experts. This is necessary because machines, of course, only operate at the syntactic level when determining semantic relatedness.[11]

Alignments are used to construct mappings between heterogeneous taxonomies, and based upon these mappings, the integration of taxonomies can be performed. Both mappings and integration will be described in subsequent sections.

As indicated so far, the existence of multiple taxonomies describing the same domain makes it necessary to introduce correspondences, alignments, and mappings. Due to the fact that multilinguality is the most prominent reason for such heterogeneity, in the following, many techniques are introduced by means of multilingual examples. These examples and techniques can usually be adopted for monolingual cases as well.

6.4.1 Structure of Taxonomy Mappings

Alignments that result from the matching calculation are essentially pairs of concepts occurring in the respective taxonomies that are attributed with confidence values. Mappings are used for the translation of the concepts of one taxonomy to concepts of another with full confidence. [BS08] [SS08] [IFLA09] are speaking of ‘target’ and ‘source’ vocabularies.

More precisely, whereas alignments relate concepts without structural restrictions, a ‘mapping’ maps a source concept to one concept or a set of concepts in the target taxonomy. Note that the mathematical

[11] Clearly, when people speak of semantics in conjunction with computers, only syntactic encodings of semantics, usually based on mathematical models, are meant rather than of semantics itself.

property of a mapping is adopted, i.e. one source concept cannot be mapped twice. These different kinds of mappings will be described in the following, with alignments being considered as the origin.

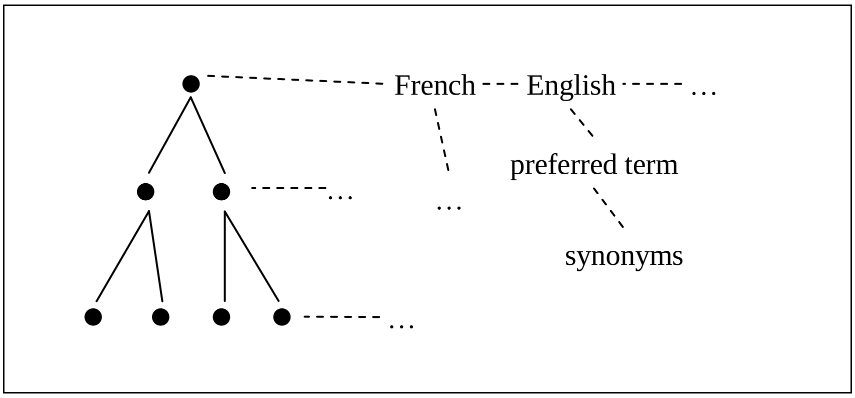


Figure 13: Unity Model for Symmetrical, Multilingual Taxonomies

The first form of mappings is symmetrical (or bijective) mapping. In this kind, every concept of the source taxonomy has an equivalent counterpart in the target. Usually, symmetrical mappings are derived from alignments that contain many equivalence correspondences (see section 6.3.1) with high confidence values. Symmetrical mappings have the form

- *$(A \rightarrow B, =)$*

where each concept *A* from the source taxonomy is mapped to exactly one concept *B* from the target taxonomy.

Moreover, symmetrical mappings are structure preserving: if a hierarchical relation exists between concepts *A* and *B* in the source taxonomy, then a hierarchical relation between the equivalents of concepts *A* and *B* also exists in the target taxonomy.

Symmetrical mappings are reflected by the unity model of [BS08]. The most prominent example of a unity model based on a symmetrical mapping is the classical multilingual vocabulary. Figure 13 shows its structure. Because of the symmetry of participating taxonomies, the structure is not duplicated. Instead, every concept is specified together with each language version that contains the preferred term and the synonym ring.

Apart from the unity model, two further architectural models exist. [BS08] proposes two different kinds of architectures. The first model provides pairwise mappings between all taxonomies, and the second is based on the well-known pivot structure (also called the ‘Backbone Model’). Details of the pivot structuring can be found in [FI08].

When different heterogeneous taxonomies have to be integrated, the symmetrical case cannot be generally assumed. In this case, the matching calculation has predominantly produced correspondences based on specialisation or overlap. According to [IFLA09] [BS08], overlap correspondences within alignments lead to one-to-one mappings (also called near-equivalence, near-match, or quasi-synonymy). Specialisation alignments lead to mappings where one concept is mapped to a set of concepts in the target taxonomy (cf. [Doe01]). This is called one-to-many mapping. For instance, let the following three correspondences (without confidence values) be given:

- *(Car, Poussette, $\supset$)*
- *(Car, Camion, $\supset$)*
- *(Car, Limousine, $\supset$)*

The above correspondences stem from two taxonomies, one English and the other French. For the English one, it is assumed that the concept granularity (see section 6.2) is weaker than in the French. Therefore, the concept *Car* is not refined, whereas in the French one all species exist.

The resulting one-to-many mapping for the case above can be represented by the form

- *(Car → (Poussette, Camion, Limousine), ∪)*

where the operator ∪ indicates that the union of the three French concepts constitutes the concept *Car* from the English taxonomy. The opposite case (many-to-one generalisation mapping) can be represented without aggregating the species:

- *(Chairs → Furniture, ⊂)*
- *(Tables → Furniture, ⊂)*

This case is equivalent to the inversion of the one-to-many mapping with operator ∪.

The second important one-to-many mapping is not connected to specialisation. Instead, this kind of mapping is necessary when a concept of the source taxonomy is only expressable in the target taxonomy by means of more than one concept. Assume that an English source taxonomy contains the concept *CarSafety*. The German target taxonomy does not provide such a concept but includes the concepts *Automobil* and *Funktionssicherheit* (functional reliability). A correct mapping for this situation is given by

- *(CarSafety → (Automobil, Funktionssicherheit), ∩)*

where ∩ is the operator expressing the intersection of the two concepts in the German taxonomy.

According to [BS08] [IFLA09] Table 6 completely summarises the kinds of mapping (right column) related to the level of equivalence (left column) of participating taxonomies that have to be mapped.

Equivalence level	Kind of mapping
Exact equivalence: two concepts have the same meaning in two taxonomies	One-to-one equivalence mapping (symmetrical construction)
Inexact equivalence: the meanings of the two concepts are not precisely identical and have overlapping extensions	One-to-one overlap mapping (quasi-synonymy)
Partial equivalence: the meaning of the concept of one taxonomy is more general that the meaning of the concept in the other or vice versa	One-to-many mapping with ∪ operator (inverse: many-to-one generalisation mapping)
Equivalence to compounds: the meaning of one concept can be expressed by a conjunction of two concepts in the other taxonomy	One-to-many mapping with ∩ operator

Table 6: Kinds of Mappings

As described earlier, mappings are directed from the source taxonomy to the target. Bidirectional mappings consist of a pair of mappings, one for each direction. Symmetrical mappings are, of course, bidirectional by definition.

6.4.2 Application of Mappings

After a mapping is created, it can be used for several applications where different taxonomies have to be integrated. The two base types of the application of mappings are:

- Merging
- Mediation

The merging of two taxonomies results in new taxonomies and optionally in a new mapping. Translation replaces concepts of the source taxonomy by concepts of the target taxonomy. Mediation translates concepts, classifications, and queries between partners using different heterogeneous taxonomies. Both basic types will be discussed in the following.

MERGING

Based on the mappings created, which are directed from the source taxonomy to the target, the target is modified or enriched by means of the source taxonomy. Vice versa, it is possible to modify the source taxonomy, too. The latter is used for creating a new mapping that only consists of equivalencies such that symmetry is reached. This symmetrical mapping ensures a proper bidirectional application necessary in mediation scenarios.

Concerning the technical application of a mapping, merging of more than two taxonomies is an iterative process where two taxonomies are merged at a time so that the result will be the input of subsequent merging operations.

Mappings are resolved during the merging process. Of course, this is a process that demands a high degree of expertise within the domain. Nevertheless, mappings can be applied in a systematic way (cf. [Doe01]). Table 7 describes the kinds of mapping based on non-exact equivalencies and their canonical resolution in the merging scenario according to [BS08] [IFLA09]. Incomplete mappings, i.e. mappings that do not relate a source concept or a concept from the target taxonomy can be resolved by introducing new concepts represented by loan terms or coined terms (see section 2.4). In this way, a symmetrical mapping can be constructed.

Mapping	**Mapping resolution**
One-to-one overlap mapping	*Target taxonomy & source taxonomy:* If concepts *a* and *b* overlap, introduce a common super concept and build a new intersection concept *c* such that *a, b, c* are disjoint. An alternative is to accept a near-match so that these concepts are treated as equivalent [IFLA09].
One-to-many mapping with ∪ operator (for many-to-one generalisation mappings the inverse resolution is applied)	*Form: (x → (a, b), ∪)* *Target taxonomy:* Introduce a new super concept *a_b* (equivalent to *x*) represented with a coined term in the target and state *a, b* as specialisations of *a_b*. (variant: use loan term in the target) S*ource taxonomy:* Introduce new concepts *x_a* and *x_b* (equivalent to *a* and *b*) in the source as specialisations of *x*. (variant: use loan terms in the source)
One-to-many mapping with ∩ operator	*Form: (x → (a, b), ∩)* *Target taxonomy:* Introduce a new concept *a_b* in the target that is represented by a coined term or loan term of *x*. In addition specialisation relationships of *a_b* to *a, b* can be

	stated. *Source taxonomy:* Introduce new concepts *a_x* and *b_x* in the source that are represented by a coined terms or loan terms. In addition specialisation relationships of *x* to *a_x*, *b_x* can be stated.

Table 7: Mapping Resolution

For a more detailed discussion and further mapping variants and examples we refer to [AW80] [BS08]. Nevertheless, an example of the resolution of a one-to-many mapping is described in the following. Assume that the English source taxonomy contains a concept (represented by the term) *Skidding* and that the German target taxonomy contains the concepts *Rutschen* and *Schleudern* together with the mapping

- *(Skidding → (Rutschen, Schleudern), ∪)*

The target taxonomy can be extended with a new concept represented by the coined term *Rutschen/Schleudern* (or directly by *Skidding* as the loan term) with its specialisations *Rutschen* and *Schleudern*. In the source taxonomy new concepts as specialisations of *Skidding* can be introduced and, furthermore, represented by loan terms or by the homographs *Skidding(sideways)* and *Skidding(forwards)*. Finally the new symmetrical mapping is given by

- *(Skidding → Rutschen/Schleudern, =), (Skidding(sideways) → Schleudern, =), (Skidding(forwards) → Rutschen, =)*

This symmetrical mapping can be used for mediation tasks.

MEDIATION

The application of mappings to translate concepts of one taxonomy to another (here also called translation, in contrast to resolution) is needed in one of the typical SEMIC.EU scenarios: mediation. In this scenario, several heterogeneous taxonomies are used by SEMIC.EU partners to classify and exchange artefacts.

Classifications have to be translated to the target taxonomy. Here it is assumed that an artefact classification is represented by a conjunction of concepts occurring in the taxonomies. A classification is translated by applying the mapping to the concepts contained in the classification.

A typical mediation scenario in the context of SEMIC.EU is a user searching for job descriptions. Each partner of this type of SEMIC.EU project can have its own taxonomy in order to construct job descriptions that are provided by repositories. Job seekers, or people in general who want to be informed about the job market, are able to search for jobs by making queries consisting of concepts from the taxonomy used in their country of residence. These will be translated to concepts from the partner’s taxonomy (possibly via a pivot taxonomy), where the query is executed. Afterwards, the classification of the job description found will be translated back to the source.

This scenario requires that concepts be translated from one taxonomy to another and back preserving their semantics. Thus at least a bidirectional mapping is required, but bidirectionality is not enough. Moreover, it is required that the mapping be complete, i.e. that all concepts can be mapped in both directions. The mapping will be applied to queries and job classifications. Several additional constraints will be given in the following discussion to ensure bidirectional semantics-preserving mappings (called constrained mappings).

Queries can be transformed directly based on the mappings shown in Table 6 with the following constraints. The first is that one-to-one overlap mappings are treated as equivalence mappings (accepting near match) or that the two participating taxonomies are merged with respect to overlaps such that a symmetric mapping can be used.

A further constraint is related to many-to-one generalisation mappings. This kind of mapping is not applicable for queries unless it is accepted that broader, but not specified, information is retrieved. For instance, if one searches for jobs that require the skill *Leadership*, the translation, according to the many-to-one generalisation mappings, could be *HumanDynamics.* So all job descriptions classified with *HumanDynamics* would be retrieved, including those classified with *SituationalAwareness* or other sub-concepts.

The last constraint is that the search engine provide constructs for union and intersection. Because these constructs are contained in the basic set of constructs for every engine (operators *AND* and *OR*), one-to-many mappings using operator $\cup$ and $\cap$ can be transformed as follows:

- *search(HumanDynamics)* and *(HumanDynamics → (SituationalAwareness, Leadership),* $\cup$*)* is transformed to *search(SituationalAwareness OR Leadership)*
- *search(ProcessSafety)* and *(ProcessSafety → (ProcessEngineering, Safety),* $\cap$*)* is transformed to *search(ProcessEngineering AND Safety)*

After retrieving the query results, i.e. artefacts (or in the present scenario job descriptions) together with their classifications[12], they have to be translated back to the query origin. Constraints for applying on-to-one mappings to classifications are the same as in the query case. The constraint for many-to-one generalisation mappings is similar to the query case and not fully applicable unless the decrease of specificity is accepted. For instance, if a job description is classified with *Leadership* that is translated to the more general concept *HumanDynamics*, the recipient will not get the specific meaning of the description.

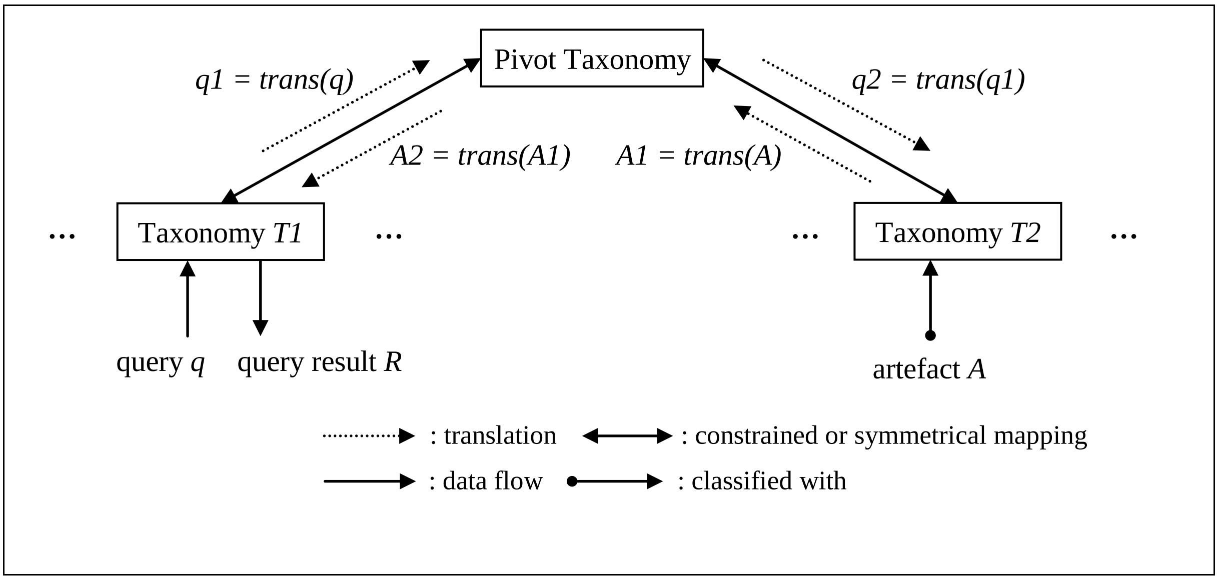


Figure 14: Mediation Scenario

[12] Interoperability (and mappings) for non-taxonomic artefacts, which can be arbitrary information models, are not considered here. Mapping of information models was discussed in [FI09]. In these guidelines, only the taxonomic part, i.e. mappings for artefact classifications is considered.

One-to-many mappings using operator $\cup$ cannot be applied because it is assumed that classifications are concept conjunctions. The latter fact, however, allows the application of one-to-many mappings using operator $\cap$. In this case, the classification is extended by the target concepts of the mapping.

Figure 14 exemplifies the typical mediation scenario. The two partners use taxonomy *T1* and *T2* for classifying their artefacts, respectively. The first partner makes a query consisting of concepts from his taxonomy in order to find artefacts classified with them, e.g. job descriptions classified with skill concepts.

A proper translation chain is guaranteed only if constrained mappings for queries and classifications are applied or if the mapping is fully symmetrical, as in the unity model [BS08]. After translating the query *q* according to the mapping definition in two steps via the pivot structure (also called backbone model in [BS08]), the translation result *q2* is used for querying taxonomy *T2,* where artefact *A* is returned. The mappings are used to translate the artefact and its classification back to the first partner, who finally gets the query result *R* wrapping *A2*. In the case of pairwise mappings between the partners, the pivot structure is not used; only the mapping between *T1* and *T2 is*.

PROCESS ASPECTS

As discussed in the [FI08], the pivot approach is to be preferred if more than three partners want to exchange information in a way that preserves semantics. Of course, this is true from the purely combinatorial viewpoint, but additional factors are relevant in conjunction with pivot and pairwise mappings. One of the key factors is the degree of heterogeneity of the taxonomies used by the partners because the amount of effort required to define a pivot taxonomy by unifying heterogeneous taxonomies depends directly on this. Another relevant factor is the overall work process for establishing the pivot architecture or the architecture based on pairwise mappings.

Assuming that *N* partners are participating in a SEMIC.EU project, the general work process is described in the following. First, the goal is to provide mappings among all *N* taxonomies. The process steps are:

1. Choose an initial taxonomy as the pivot taxonomy.
2. Iteratively modify the pivot taxonomy by matching and mapping it to the *N* participating taxonomies and, subsequently, perform a merge where only the pivot taxonomy is extended.
3. For each participating taxonomy, based on the corresponding mapping from step 2, perform a matching and construct a bidirectional mapping between the partner taxonomy and the pivot taxonomy

In step 1, it is important to choose a taxonomy that has a high degree of conceptual and structural granularity (cf. section 6.2) in order to cover most of the concepts used in the participating partner taxonomies. This initial pivot taxonomy could be one of the partner taxonomies or a taxonomy generally accepted by the community (the pivot taxonomy has the same status as the so-called ‘dominant language’ from [AW80]).

Step 2 seeks to align the initial pivot taxonomy to all partner taxonomies stepwise. Matching and mapping techniques described above are performed during step 2. Based on these techniques, a merge is performed in order to resolve semantic heterogeneity within the pivot taxonomy as much as possible. Within step 3, the final mappings between each partner and the pivot taxonomy are constructed. This step can rely on existing alignments and mappings from step 2.

The construction process based on the architecture of pairwise mappings is, of course, less manifold but has a higher number of iterations due to the combinatorial size mentioned above. This case is comparable with step 3, but all of the partner taxonomies have to be considered rather than the pivot taxonomy. Specifically, all individual partners have to perform a matching and have to construct a mapping from their own taxonomy to every other partner taxonomy.

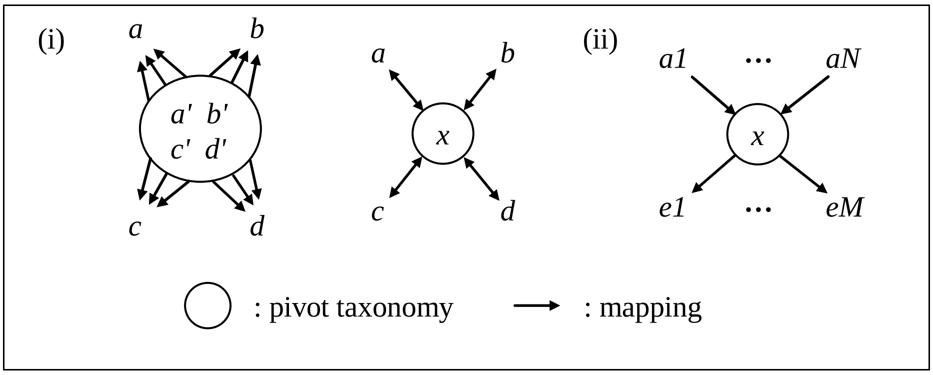


Figure 15: Pivot Structuring

In order to understand the relationship between pivot and pairwise structuring, it is useful to identify the extremes of the pivot approach. In one extreme, the pairwise structure is fully lifted to the pivot structure, whereas the latter case conforms fully to the symmetric model (unity model) explained above. Figure 15 (i) shows both extremes, assuming four partner taxonomies. The extreme on the left side is distinguished by the fact that every partner taxonomy is represented one-to-one in the pivot taxonomy, i.e., *a, b, c,*and *d* are represented by their clones *a', b', c', d'*. This yields a situation where the extent of the mapping to the pivot taxonomy is the same as in the case of the pairwise approach, e.g. *b', c', d'* has to be mapped to *a*. But because the pivot taxonomy has to be managed in addition, the effort for the pivot process is higher.

The other extreme on the right side of (i) uses only symmetrical one-to-one mappings, as is done in the case of classical multilingual thesauri (cf. 6.4.1). In contrast to the left extreme, here the classical advantage of the pivot structure exists because available mappings can be fully reused. Reuse of existing mappings is responsible for the efficiency of the pivot architecture.

Figure 15 (ii) exemplifies the reuse of existing mappings in contrast to pair-wise mapping definitions. Here, *N+M* taxonomies, where at least *N* taxonomies have parts that are in symmetrical relationship, are assumed. Based on the pivot approach, concept *x* is mapped to concept expressions (e.g. one-to-one or with operator $\cap$ or $\cup$) over concepts from each of the *M* taxonomies. If the concepts *x1,…,xN* from the *N* taxonomies are mapped to the concept *x* from the pivot taxonomy, all of the mappings outgoing from *x* are reused, so that we have *N+M* mappings in total. Obviously, based on the pairwise architecture, *N*M* mappings would be needed.

In a real-world application, the possible pivot structure is located between the two extremes, i.e. many but not all parts of the partner taxonomies are in symmetrical relationship. Then reuse of corresponding mappings can be done such that the overhead due to the construction of the pivot taxonomy in process step 1 and 2 is justified. The point at which the degree of reuse is too small to be efficient depends on the project and particularly on the degree of heterogeneity. In general, also motivated by the fact that working with the pivot approach makes it possible to get a unified model of the domain, the reuse of already-introduced mappings is expected to be high enough to justify the pivot approach.

7. CONCLUSIONS

Taxonomies are widely used to structure and classify all kinds of information in a hierarchical system. As almost anything can be classified according to some taxonomic scheme, taxonomies also have a huge impact on the interpretation of the information classified, in other words on semantic interoperability.

This document has discussed the whole life cycle of taxonomies with a view toward their technical, methodological, conceptual, and organisational aspects. Based on a detailed definition of taxonomies and their constituents, a list of usage scenarios was provided in order to span the scope of the problems encountered, covering the construction and maintenance of a single taxonomy, preferably by reusing an existing one, as well as syntactic and semantic integration of different taxonomies.

The crucial issues that have to be considered in order to have a good taxonomy are:

- Applying a structured process in analogy to established process models in software engineering. Such a structured process covers all the relevant steps, from the determination of the scope and requirements of a taxonomy to its management and maintenance.
- Considering the reuse of an existing taxonomy, e.g. from the SEMIC.EU platform, as the reuse of an existing taxonomy has the potential for cost savings and a high degree of quality, in particular, when it has passed the SEMIC.EU quality process.
- If no appropriate taxonomy for reuse could be found, a collaboration and community process on the SEMIC.EU platform should be started to get in contact with similar projects and to involve the interested community in the quality improvement of the taxonomy.
- Deciding on the right type of taxonomy.
    - A monohierarchical taxonomy is appropriate for classifying terms unambiguously, as required in statistical and biological applications.
    - A polyhierarchical taxonomy is appropriate when a classification has to be established with regard to different criteria in parallel, e.g. in medical applications.
    - Generally speaking, faceted classification is the best solution if applicable. A necessary condition is that all objects within the domain share the same properties expressed by the respective facets.
    - Associative relationships may be useful for enhancing the expressivity of a taxonomy, but it has to be taken into account that this implies a higher degree of complexity in construction and maintenance.
- Choosing a hierarchical structuring based on the following criteria.
    - Appropriate classification criteria with regard to the domain and purpose of the taxonomy
    - Appropriate specificity/granularity

Machine readability is a sine qua non condition for electronic data exchange. In particular, taxonomies themselves have to be machine readable. This requires a representation language for taxonomies. Preferred candidates are the standards

- SKOS for representing explicitly defined taxonomies
- OWL for taxonomies defined by means of terminological expressions
- RIF for F-logic-based taxonomies.

An exchange of taxonomies represented by different languages requires their syntactic integration. This document provides a reference architecture that allows for exchanging customised languages as well as standard languages between and among SEMIC.EU project partners.

Semantic integration is one of the main issues in the pan-European context when different communication partners employ different heterogeneous taxonomies. Different kinds of heterogeneity (namely, terminological and conceptual) have to be dealt with and resolved, if possible. These guidelines present the means for detecting heterogeneity (matching). Furthermore, it identifies when to use which matching technique, depending on the expressivity of the taxonomy languages involved.

Different kinds of mapping between taxonomies can be used in order to resolve heterogeneity and to translate concepts between them. Finally, these techniques can be applied in order to

- merge existing taxonomies for creating new ones
- build mediators for the purpose of translating queries and artefacts classified according to decentralised taxonomies.

In this document hierarchical knowledge representation has been identified as an integral component of current and future European eGovernment practice. Taxonomies are one of the fundamental instruments in the creation of semantic interoperability in general. From this perspective, they can be seen as both a subject-matter of interoperability and a tool used to facilitate the interoperation of diverse systems.

Taxonomies are a primary phenotype of the structuring of complex knowledge. The main challenge faced by everybody involved in the classification of social, political and administrative real-world knowledge as well as in the strife for seamless pan-European data exchange is how to systematise the factual knowledge and practical procedural interdependencies.

This document highlights the great potential of taxonomic knowledge representations for pan-European eGovernment projects. It elaborates on the benefits of such strategies as reuse and harmonisation and, thus demonstrates the essential importance of taxonomies in international, transnational and pan-European semantic interoperability.

Appendix A REFERENCES AND LITERATURE

[AELP09] d'Aquin, M., Euzenat, J., Le Duc, Ch., and Palma, R. Cupboard – Supporting Ontology Reuse by Combining a Semantic Web Gateway, Ontology Registry and Open Ratings Systems, NeOn: Lifecycle Support for Networked Ontologies, Integrated Project (IST-2005-027595). 2009.

[AGR09] AGROVOC Thesaurus – Agricultural Information Management Standards. http://www.fao.org/agrovoc/ (last visited 02.09.2009)

[AGY05] Avesani, P., Giunchiglia, F., and Yatskevich, M. A Large Scale Taxonomy Mapping Evaluation. International Semantic Web Conference. 2005.

[ANS05] ANSI/NISO Z39.19 – Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. National Information Standards Organisation. 2005.

[AW80] Austin, D., and Waters, J. Guidelines for the Establishment and Development of Multilingual Thesauri, Revised Text. United Nations Educational, Scientific and Cultural Organization. 1980.

[Bar03] Barnes, J. (ed.). Porphyry Introduction (Clarendon Later Ancient Philosophers). Oxford University Press. 2003.

[BCC06] Brusa, G., Caliusco, M., and Chiotti, O. Building Ontology in Public Administration: A Case Study. In: Proceedings of the First International Workshop on Applications and Business Aspects of the Semantic Web (SEBIZ). 2006.

[BCM+03] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., and Patel-Schneider, P.F. The Description Logic Handbook: Theory, Implementation, Applications. Cambridge University Press. 2003.

[BEF+04] Bouquet, P., Euzenat, J., Franconi, E., Serafini, L., Stamou, G., and Tessaris, S. D2.2.1: Specification of a common framework for characterizing alignment. Technical report, NoE Knowledge Web project. 2004.

[BHH+04] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., and Stein, L.A. OWL Web Ontology Language Reference. W3C Recommendation. W3C Consortium. 2004.

[BHL01] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. In: Scientific American Magazine. 2001.

[BL04] Brachman, R., and Levesque, H. Knowledge Representation and Reasoning. Morgan Kaufmann Series in Artificial Intelligence. 2004.

[BLN86] Batini, C., Lenzerini, M., and Navathe, S.B. A comparative analysis of methodologies for database schema integration. In: ACM Computing Surveys. 1986.

[Bro04] Broughton, V. Essential Classification. London: Neal-Schuman Publishers Inc. 2004.

[BS05] British Standard BS 8723-2:2005. Structured vocabularies for information retrieval – Guide. Part 2: Thesauri. 2005.

[BS08] British Standard BS 8723-4:2007. Structured vocabularies for information retrieval – Guide. Part 4: Interoperability between vocabularies. Incorporating Corrigendum No. 1. 2008.

[Cho06] Choksy, C.E.B. 8 Steps to develop a taxonomy. In: Information Management Journal 40(6). 2006.

[CL07] Common Logic Standard. http://common-logic.org/ (last visited 06.07.2009).

[CRF03] Cohen, W., Ravikumar, P., and Fienberg, S. A comparison of string metrics for matching names and records. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining. 2003.

[Dex01] Dextre Clarke, S.G. Thesaural relationships. In: Bean, C.A., and Green, R. (eds.): Relationships in the Organization of Knowledge. Boston: Kluwer. 2001.

[Doe01] Doerr, M. Semantic problems of thesaurus mapping. In: Journal of Digital Information. 2001.

[EPR09] ePractice. http://www.epractice.eu (last visited 02.09.2009).

[ES07] Euzenat, J., and Shvaiko, P. Ontology Matching. Berlin/Heidelberg: Springer-Verlag. 2007.

[Fer99] Fernández López, M. Overview of Methodologies for Building Ontologies. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)-99 workshop on Ontologies and Problem-Solving Methods (KRR5). 1999.

[FI08] Fraunhofer ISST and]init[. Study on Multilingualism. European Commission – IDABC. 2008.

[FI09] Fraunhofer ISST and]init[. Study on Methodology. European Commission – IDABC. 2009.

[Gau05] Gaus, W. Dokumentations- und Ordnungslehre. Theorie und Praxis des Information Retrieval. Berlin/Heidelberg: Springer-Verlag. 2005.

[GEN09] Gene Ontology. http://www.geneontology.org/ (last visited 03.09.2009).

[GSY04] Giunchiglia, F., Shvaiko, P., and Yatskevich, M. S-Match: an algorithm and an implementation of semantic matching. In: Proceedings of the European Semantic Web Symposium. 2004.

[HP05] Hjørland, B., and Petersen, K.N. A substantive theory of classification for information retrieval. In: Journal of Documentation 61. 2005.

[ICD09] International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/ (last visited 02.09.2009).

[IFLA09] Guidelines for Multilingual Thesauri. International Federation of Library Associations and Institutions IFLA Professional Reports, No. 115. 2009.

[Joh04] Johansson, I. On the transitivity of the parthood relation. In: Hochberg, J., and Mulligan, H. (eds.): Relations and Predicates. Frankfurt: Ontos. 2004.

[Jun08] Jung, J. Taxonomy alignment for interoperability between heterogeneous virtual organizations. In: Expert Systems with Applications, vol. 34. 2008.

[KLW95] Kifer, M., Lausen, G., and Wu, J. Logical Foundations of Object-Oriented and Frame-Based Languages. In: Journal of the ACM. 1995.

[KN06] Khoo, C., and Na, J.C. Semantic Relations in Information Science. In: Annual Review of Information Science and Technology, 40. 2006.

[Kom92] Komatsu, L.K. Recent view of conceptual structure. In: Psychological Bulletin 112(3). 1992.

[LN07] Leser, U., and Naumann, F. Informationsintegration. Heidelberg: dpunkt.verlag, 2007.

[MAGP08] Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. Modelling Multilinguality in Ontologies. International Conference on Computational Linguistics (COLING). 2008.

[MB09] Miles, A., and Bechhofer, S. SKOS Simple Knowledge Organisation System Reference. W3C Consortium. 2009.

[MSH08] Medical Subject Headings. Bethesda, MD: U.S. National Library of Medicine. 2008.

[NAC08] Statistical classification of economic activites in the European Community. Rev. 2. Eurostat. 2008.

[NM01a] Noy, N.F., and McGuiness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology. 2001.

[NM01b] Noy, N.F., and Musen, M. Anchor-PROMPT: using non-local context for semantic matching. In: Proceedings of the workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence. 2001.

[NMVI09] Nickerson, R., Muntermann, J., Varshney, U., and Isaac, H. Taxonomy Development in Information Systems: Developing a Taxonomy of Mobile Applications. In: Proceedings of the 17th European Conference on Information Systems (ECIS). Verona, Italy. 2009.

[Noy04] Noy, N.F. Semantic integration: a survey of ontology-based approaches. SIGMOD. 2004.

[NP01] Niles, I., and Pease, A. Towards a standard upper ontology. In: Proceedings of the international Conference on Formal ontology in information Systems. 2001.

[PC96] Park, Y.C., and Choi, K.S. Automatic thesaurus construction using Bayesian networks. In: Information Processing & Management 32. 1996.

[PF07] Piechocki, M., and Felden, C. XBRL Taxonomy Engineering. Definition of XBRL Taxonomy Development Process Model. In: Proceedings of the 15th European Conference on Information Systems (ECIS2007). 2007.

[PM05] Paslaru Bontas, E., and Mochol, M. A Cost Model for Ontology Engineering. Technical Report B-05-03, Free University Berlin. 2005.

[PMT05] Paslaru Bontas, E., Mochol, M., and Tolksdorf, R. Case Studies on Ontology Reuse. In: Proceedings of the 5th International Conference on Knowledge Management IKNOW05. 2005.

[Ras05] Raschen, B. A resilient, evolving resource – How to create a taxonomy. In: Business Information Review (22). 2005.

[RIF09] RIF Working Group. http://www.w3.org/2005/rules/wiki/RIF_Working_Group (last visited 06.07.2009).

[RLG04] Rosati, L., Lai, M., and Gnoli, C. Faceted Classification for Public Administration. In: Proceedings of Semantic Web Applications and Perspectives (SWAP). 2004.

[RM06] Rosenfeld, R., and Morville, P. Information architecture for the World Wide Web. O'Reilly Media. 2006.

[Ros78] Rosch, E. Principles of Categorization. In: Lloyd, R.R., and Lloyd, B.B. (eds.): Cognition and Categorization. Hillside, NJ: Lawrence Erlbaum Publishers. 1978.

[SB05] Schneider, J.W., and Borlund, P. A bibliometric-based semi-automatic approach to identification of candidate thesaurus terms: Parsing and filtering of noun phrases from citation contexts. In: Lecture Notes in Computer Science 3507. 2005.

[SCK+05] Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., and Rosse, C. Relations in biomedical ontologies. In: Genome Biology 6(5), Art. R46. 2005.

[SE05] Shvaiko, P., and Euzenat, J. A Survey of Schema-based Matching Approaches. In: Journal on Data Semantics IV. 2005.

[SHS00] Steinberger, R., Hagman, J., and Scheer, S. Using Thesauri for Automatic Indexing and for the Visualisation of Multilingual Document Collections. In: Proceedings of the workshop on Ontologies and lexical knowledge bases. 2000.

[SKK04] Smith, B., Köhler, J., and Kumar, A., On the Application of Formal Principles to Life Science Data: a Case Study in the Gene Ontology. In: International Workshop on Data Integration in the Life Sciences (DILS). Berlin/Heidelberg: Springer-Verlag. 2004.

[SL90] Sheth, A., and Larson, J. Federated database systems for managing distributed, heterogeneous, and autonomous databases. In: ACM Computing Surveys. 1990.

[SS08] Stock, W., and Stock, M. Wissensrepräsentation. München: Oldenbourg Verlag. 2008.

[SWG02] Shasha, D., Wang, J.T.L., and Giugno, R. Algorithmics and applications of tree and graph searching. In: Proceedings of the Symposium on Principles of Database Systems. 2002.

[TBL09] Thau, D., Bowers, S., and Ludäscher, B. CleanTax: A Framework for Reasoning about Taxonomies. Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium. Technical Report SS-09-02. 2009.

[TSCA02] Tzitzikas, Y., Spyratos, N., Constantopoulos, P., and Analyti, A. Extended faceted taxonomies for Web catalogs. In: Proceedings of the Third International Conference on Web Information Systems Engineering (WISE). 2002.

[UG96] Uschold, M., and Gruninger, M. Ontologies: Principles, Methods and Applications. In: Knowledge Engineering Review 11(2). 1996.

[UJ07] Uddin, M.N., and Janecek, P. The implementation of faceted classification in web site searching and browsing. In: Online Information Review 31(2). 2007.

[VS09] Vrandecic, D., and Sure, Y. Ontology Repositories and Content Evaluation. Deliverable D1.2.10v2 NEON Project, http://www.aifb.uni-karlsruhe.de/WBS/dvr/publications/kwebd1210.pdf (last visited 02.07.2009)

[WB08] Whittaker, M., and Breininger, K. Taxonomy Development for Knowledge Management. World Library and Information Congress. 74th IFLA General Conference and Council. 2008.

[WCH87] Winston, M.E., Chaffin, R., and Herrmann, D. A taxonomy of part-whole relations. In: Cognitive Science 11. 1987.

[Wit53] Wittgenstein L. Philosophische Untersuchungen. Frankfurt: Suhrkamp. 1953.

[Wol97] Wolters, Ch. GOS Thesaurus-Handbuch. Institut für Museumskunde Staatliche Museen zu Berlin Preußischer Kulturbesitz. Technical Report TR 97-19. 1997.

[WSML08] The Web Service Modeling Language WSML. http://www.wsmo.org/wsml/wsml-syntax (last visited 06.07.2009).

Appendix B SUMMARY OF PRINCIPLES, METHODS, AND GOOD PRACTICES

The following table summarises the findings of this document. The overall structure of the table reflects the structure of the construction process of a taxonomy. The first row (“principal use cases”) designates the situation in which a rule, technique, or good practice is to be applied. In the second row (“rule/technique/good practice”), a short version of the corresponding rule, technique, or good practice is given.

Principal use case	**Rule/Technique/Good Practice**
Determine requirements	
Basic questions to be answered	Questions to be answered concern scope, intended use, user groups and maintainers of the planned taxonomy.
Consider reusing existing taxonomies	
In general	In order to avoid "duplicating" an existing taxonomy, reusing a taxonomy of the same or an overlapping domain should be a goal.
In general	A cost model should be applied in order to conduct a cost/benefit analysis.
In particular, under the following conditions: There are well-established taxonomies available that are applied in contexts quite similar to the one in which the planned taxonomy is to be applied. One single reuse candidate covers most of the domain the target taxonomy has to cover, i.e. is not necessary to merge different taxonomies. The identified reuse candidate is implemented in the same taxonomy language as the planned taxonomy is to be implemented in, or is at least implemented in a standard taxonomy language, not in a proprietary one.	Reuse should be considered as under these conditions cost savings and quality improvement are likely.
Identify Concepts and Determine Relationships	
Homographs	Homographs must be disambiguated by qualifiers.
Additional information for terms	Scope notes, definitions, or history notes may be attached to terms in order to add information about usage, distinction from other terms, meaning, and development over time.
Terms – unclear meaning	Definitions should be added to a term when the meaning of a term is not clear for all user groups.
Form	Cf. [ANS05] [BS05].
Sources for term identification	Reference sources, textbooks, encyclopaedias, etc. about the domain. Text-based content objects. Vocabularies different user groups are used to.
Sources for term identification include mainly reference sources, textbooks, encyclopaedias, etc. about the domain	Apply top-down approach for selection of terms.
Sources for term identification include mainly text-	Apply bottom-up approach for selection of

Principal use case	**Rule/Technique/Good Practice**
based content objects	terms.
Identification of terms for complex taxonomies, huge number of text-based sources	Machine assistance should be used to support term identification
Synonyms	True synonyms should be included in the taxonomy. Among synonyms, one must be designated as the preferred term.
Near synonyms	Near synonyms should not be treated as synonyms within the core area of a domain.
Generality/specificity of terms	Specific within the core area of the subject field, depending on application. Trade-off between economy and information content has to be considered.
Completeness of hierarchy	Every term in a taxonomy has to have a hierarchical relationship to at least one other term.
Definition of hierarchy – choice of classification criteria	The hierarchy should be determined by means of properties that are definitely appropriate for the distinction of terms and with respect to the purpose of the taxonomy. Due to the limitations of characterising terms by common sets of properties, terms may be classified according to “family resemblance”. Subclasses of a class usually have additional properties, restrictions, or different relationships compared to the superclass.
Disjunctiveness of terms – required by some applications for unambiguous classification	Structure of taxonomy has to be monohierarchical or faceted.
Monohierarchy versus Polyhierarchy	Trade-off between less complexity of monohierarchicy and flexibility of polyhierarchy should be considered.
Polyhierarchy . flexibility versus complexity	Trade-off between flexibility and complexity/understandability should be considered, i.e. classifying criteria should be restricted to necessary ones with regard to the purpose of the taxonomy.
Use of Faceted classification – common dimensions for all objects within the domain	Always, when possible, i.e. common dimensions for all objects within the domain. In particular, for ▪ New and emerging fields with incomplete domain knowledge ▪ Interdisciplinary areas with more than one perspective to look at content objects or

Principal use case	Rule/Technique/Good Practice
	need for combination of concepts ▪ Taxonomies with multiple hierarchies but unclear boundaries ▪ Classifying electronic documents and content objects where location and collocation is not of importance
Whole-part relationship – type(s)	For consistency and information retrieval purposes, transitivity should be granted, i.e. the use of different types of whole-part relationships within a single taxonomy should be avoided as far as possible.
Adding of associative relationships	Required if advanced expressivity is indispensable. Otherwise, trade-off between expressivity on the one hand and complexity and interoperability on the other has to be considered.
Level of generality/specificity of siblings	All the siblings in a hierarchy should have the same level of generality.
Minimum number of narrower terms of a broader term	If a class has only one direct subclass, the taxonomy is not complete or modelling may not be appropriate.
Maximum number of narrower terms of a broader terms	If there are more than a dozen subclasses for a given class it may be considered to add intermediate categories.
Review with Users and Experts; Refine Taxonomy	
What has to be evaluated	Usability, Structure, Completeness, and appropriateness of granularity
How can usability be evaluated	Let several users index the same content objects and define searches, compare results.
Implement and Test	
Which language should be chosen for implementing the taxonomy	Prefer one of the standards SKOS, OWL or RIF.
Manage and Maintain	
Candidate terms for addition	Proposals from indexers and users. Evaluation of search logs.
Candidate terms for modification and deletion	Overused and very infrequently used terms should be considered as candidates for modification or deletion.
Consequences of changes	Re-indexing of content objects may have to be considered.

Principal use case	**Rule/Technique/Good Practice**
Consequences of modification and deletion of terms	For modified and deleted terms, consideration should be given to keeping them in the taxonomy for retrieval purposes.
Supervision of changes	Changes should be supervised by a control board.
Frequency of changes	Changes should be subject of version control and release management.
Dissemination of new versions	Preferably whole new taxonomies should be distributed. Target systems linked to other data, as a minimum, may require reports or files containing the changes
Multilingual Taxonomies	
Choice of organisation structure	A semi-centralised structure should be chosen
Choice of construction approach	One of the three approaches ▪ From scratch ▪ Translation of existing taxonomy ▪ Merging of existing taxomies has to be chosen
Choice of coordination among languages	Feedback should be enabled, i.e. changing the form or structure in one language in order to achieve an easier or more useful solution in another language
Symmetrical versus non-symmetrical taxonomy	A symmetrical taxonomy should be preferred in order to avoid mapping.
Decisions to be taken for non-symmetrical taxonomies	Primary language may have to be determined.
Missing appropriate term in one language	Use of a loan term or building a coined term should be considered.
Using a non-symmetrical taxonomy	Mapping techniques have to be applied.
Syntactic Integration	
Using the same taxonomy language	No mapping has to be applied. If the project is expected to involve other languages, mapping to the corresponding standard should be provided.
Using taxonomy languages within the same expressivity group	Only mappings within one group have to be applied. This is to be done via the standard related to the group.
Using taxonomy languages from different expressivity groups	Mappings to the standard of the group followed by mappings between standards and again

Principal use case	**Rule/Technique/Good Practice**
	mappings to the group have to be applied.
Semantic Integration	
Pure taxonomies	The following approaches are adequate for taxonomy matching: string-based, language-based, linguistic resources, taxonomic structure, and repositories of structures.
Taxonomies with associative named relations between concepts	Graph-based matching techniques are applied.
Taxonomies with schema elements	Constraint-based techniques are applied.
Taxonomies with logical rules	Model-based techniques will be adequate.
Exact equivalence between concepts of a taxonomy	One-to-one mapping (symmetrical construction) should be constructed.
Near-equivalence between concepts of a taxonomy	One-to-one overlap mapping accepting near match (quasi-synonymy) should be constructed.
Partial equivalence between concepts of a taxonomy	One-to-many mapping with ∪ operator or many-to-one generalisation mapping should be constructed.
Equivalence to compound taxonomy concepts	One-to-many mapping with ∩ operator should be constructed.