Semantic Interoperability Centre Europe

Study on multilingualism

Issue: Version 1.0

Date: 2008-12-09

Authors: Fraunhofer ISST /]init[

Document Change History

Date	Version	Author	Change Details
2008-11-10	0.1	AB, JG	Initial Structure
2008-12-03	0.2	AB, JG, HA	Initial Draft
2008-12-07	0.3	AB, JG, HA	Consolidated Draft
2008-12-09	0.4	SB	QA, Finalisation
2008-12-09	1.0	Stephan Meyer, Renke Fahl-Spiewack	QA

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:

- SEMIC.EU will provide access to interoperability assets that have been developed in previous governmental projects.
- A clearing process will safeguard certain rules and standards to assure the quality of published assets.
- Community features will be available on the platform, e.g. a forum to discuss best practices for the use of assets.
- SEMIC.EU will invite stakeholders to seminars and workshops that are related to its activities.
- 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 the publication of sophisticated Semantic Interoperability Assets for the pan-European data exchange between public administrations. Semantic interoperability implies that the meaning for the sender and the receiver of a message is the same or at least compatible. With respect to one of the objectives from the EU Commissioner for Multilingualism, in a pan-European context the Member States will provide their data in their own language, too: *"Give citizens access to European Union legislation in their own languages"*.

Therefore, semantic interoperability implies that data exchanged in a pan-European context needs to be translated to the receivers' own language and individual data structure. Hence, the data exchanged as part of a pan-European communication has to be mapped from the originating data format of the sender to the data format of the receiver. Furthermore, the terminology and vocabulary has to be translated from the source to the target language. The data exchange has to be semantics-preserving, i.e. sender and receiver must have a common and ideally identical understanding of the meaning of the data in all languages involved, any incompleteness and ambiguity have to be adequately addressed.

This study presents an initial approach for dealing with multilingualism in the context of SEMIC.EU, i.e. how multilingualism should be incorporated in Semantic Interoperability Assets, how SEMIC.EU's platform functionality should be enhanced to better address multilingualism, and how best to interconnect pan-European federated applications. Based on this input the SEMIC.EU community and all relevant stakeholders are invited to contribute to the further improvement of SEMIC.EU's platform, concepts and methodology.

Mapping between different languages should be performed by using pivot mapping and appropriate mapping languages. All data exchanged as well as the defining artefacts within a Semantic Interoperability Asset should be available in the pivot language accepted by all partners. Usually English is used as the pivot language in the context of the European Union. It is highly advisable to widely use the pivot language, e.g. for identifiers, in technical artefacts like XML schemata, etc. The pivot mapping reduces the number of mappings.

This approach exploits two elementary mapping techniques. Schema mapping, on the one hand, can be used for structural changes and is a syntactic method to solve semantic issues. The usage of controlled vocabularies, on the other hand, requires more sophisticated techniques, such as taxonomies, multilingual thesauri, or ontologies. These techniques offer powerful means to translate terms on a semantic level superior to pure machine translations.

In addition, it should be investigated whether the technique of semantic tagging and semantic statements, i.e. augmenting data by topics or references supplying additional helpful information, is a valuable contribution to achieve semantic interoperability in the presence of multilingualism. Semantic tagging is an innovative and promising approach currently subject to various research activities.

Concerning the SEMIC.EU platform, besides providing the option to choose between different languages for the frontend, in particular, a multilingual search should be implemented. This search mechanism should support defining a search request in the user's local language and deliver search results that are not available in either the user's local language or the pivot language.

Finally, the study outlines a defined process for the step-wise rollout of pan-European networked applications based on an initial bilateral connection that is extended to incorporate further participants and related connections. This process proposes a mentoring model, i.e. a participant already connected to the network serves as a coach in connecting a further participant to the network.

1. INTRODUCTION

The main goal of SEMIC.EU is to publish high-quality Interoperability Assets for a broad range of users across Europe. According to Leonard Orban, the current EU Commissioner for Multilingualism, one of his organisation's objectives is: "*The European Commission needs to deliver results for citizens, and we need to communicate with you in a language you can understand. Promoting multilingualism is an excellent way to bring European citizens closer to each other. To give you access to information and to contributing your views*". Therefore, pan-European applications should support multilingualism and hence, Semantic Interoperability Assets should support multilingualism.

If the exchange of data from one organisation to another implies transforming the data from one language to another, semantic interoperability requires that there is common and ideally identical understanding of the meaning of the data in both languages. However, semantic interoperability in a multilingual context is not just a matter of different languages. Moreover, it has to consider different concepts, legal systems and cultures in a pan-European context.

In order to successfully achieve semantic interoperability in SEMIC.EU's multilingual context, appropriate mechanisms have to be supplied to map data between different languages. Moreover, Semantic Interoperability Assets have to cope with multilingualism in an appropriate manner, e.g. need to provide mapping mechanisms, provide the documentation in different languages, etc.

1.1. The Purpose of this Document

This study is aimed at presenting an initial approach on how to address multilingualism in developing Semantic Interoperability Assets, in providing an appropriate SEMIC.EU frontend including multilingual search, and in rolling out pan-European federated applications. This approach shall serve as a base for further discussion involving all relevant stakeholders in order to improve the methodology for developing Semantic Interoperability Assets, improve the SEMIC.EU platform and improve the process of establishing pan-European federated and communicating applications.

1.2. The Structure of the Document

The document starts with introducing related fields of work in the second section. In the third section, the conceptual base for implementing multilingualism in Semantic Interoperability Assets is introduced. This base comprises pivot mapping and appropriate mapping definition techniques. The fourth section describes how to implement the mechanisms of the conceptual base in Semantic Interoperability Assets, how to enhance the platform by providing support for different languages and, in particular, how to implement multilingual search, and finally it describes a process for rolling out a federated pan-European application. The fifth section summarises the proposals developed in section four and derives corresponding

recommendations for the further improvement of the SEMIC.EU platform, related concepts and methodologies, and additional promoting activities.

2. RELATED WORK CONCERNING MULTILINGUALISM IN SEMIC.EU

The fundamental objective of semantic interoperability in the context of SEMIC.EU is the reliable exchange of information in an international scope. This means that when messages are sent from organisation A to organisation B, it is essential that both have an identical understanding of the meaning of the transferred data.

If the data is incomplete or has been modified, this should be obvious to both communication partners. In Figure 1, the important layers of interoperability in the context of SEMIC.EU are shown [1].

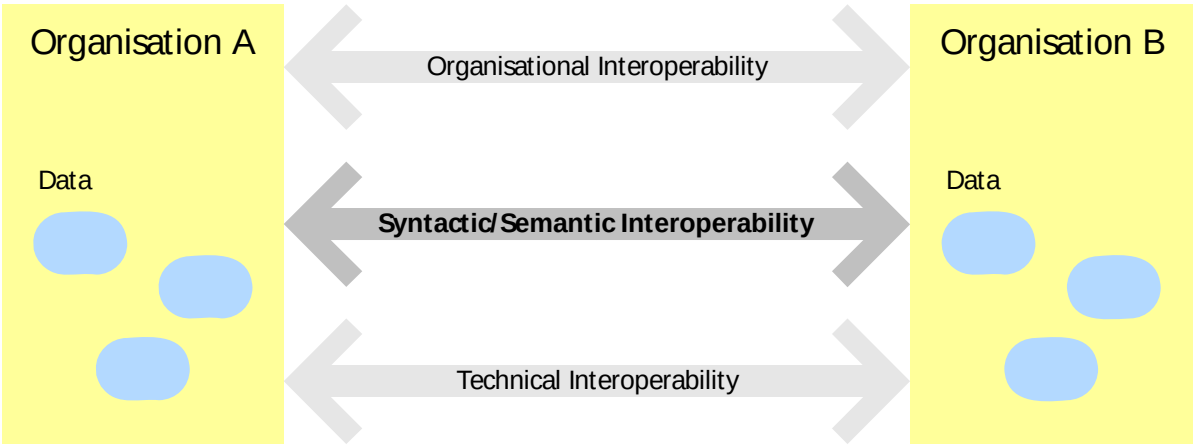

Figure 1: Layers of Interoperability

Technical interoperability is achieved by using standard network protocols like TCP/IP and transport protocols like electronic mail or web services. Organisational interoperability means that the communication partners have defined the underlying business cases and signed the required service contracts. Technical and organisational interoperability are out of scope of SEMIC.EU, which mainly focuses on the data to be exchanged. This means that mutual understanding of the application data to be exchanged is of major importance. This includes both syntactic and semantic aspects that can't be strictly isolated. In particular, some semantic problems can be avoided by changing the syntax of the data structures to be transferred, e.g., instead of transferring the name in the form "*Chan Jackie*", the first name "*Jackie*" and last name "*Chan*" can be transferred as separate information elements.

Multilingualism is a special aspect of semantic interoperability. It refers to the objective that the meaning of data should be preserved even when data is exchanged between partners from different countries using different languages and having different cultures, legal systems, and economic systems.

The European Commissioner for Multilingualism defines the term "multilingualism" as follows: "*Multilingualism refers to both a person's ability to use several languages and the co-existence of different language communities in one geographical area*"[2]. A definition more closely related to the functions of SEMIC.EU headed "*Multilingualism in computing*" can be found in Wikipedia: "*In computing, software is said to be multilingual when the user interface language can be switched. Translating user interface is usually part of the software localization process, which also includes other adaptations such as the conversion of units and dates*".

Synonyms used for multilingual are “cross-lingual” and “cross-language”, e.g. “cross-language information retrieval” (CLIR).

2.1. Schema Mapping

Before we look in detail at the multilingual issues related to semantic interoperability, we have to identify the cases where mapping of data is required. Please see the example in Figure 2, which shows two data instances and the required semantic mapping.

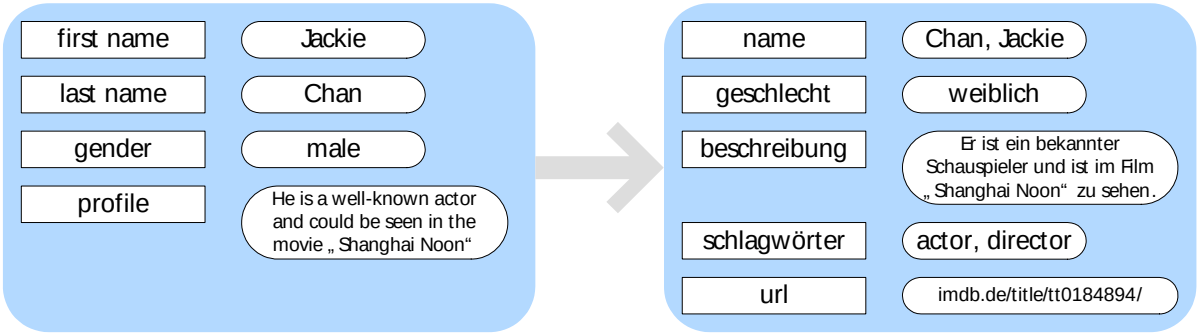

Figure 2: Example of a Data Instance

The individual’s personal description contains four pairs of values. In practical applications, these can be structured horizontally as well as in a hierarchy. The example demonstrates the following issues:

- Two properties can be combined to form a new property.
- In the reverse case, a property can be split into two multiple properties. It should be noticed that the rules governing how to split up a name can be rather complex and not unambiguous, e.g. in some Asian languages, the last name precedes the first name.
- The property must be renamed.
- There are intensional values that should not be changed, like the person’s first name.
- There can be intensional values that will change in a multilingual mapping, e.g. the name of a city like “Brussels” is “Brüssel” in German.
- There are extensional values, like the terms for the various genders, which can be enumerated. These values typically form a controlled vocabulary. Nevertheless, mapping these values to the localized term might be required.
- There might be portions of a text that should be translated.
- It might make sense not to translate portions of a text so they are available if required. Instead, the mapped data instance is enhanced by additional information, e.g., adding keywords from a controlled vocabulary that describes the original content.
- Another variant of semantic tagging might involve linking additional resources, e.g., original sources located on a website.

The original work about semantic interoperability arises in the area of database technologies. For federated systems in particular, the basics for schema mapping has been studied [3]. These works analyse the possibilities for mapping properties between two schemata most effectively. These include renaming of properties as well as decomposing and translating property values. This kind of mapping can be implemented by using well-known

technologies like Extensible Stylesheet Language Transformations (XSLT), XQuery, and Regular Expressions. These technologies use methods that are more syntactically oriented.

When it comes to multilingualism, the renaming and restructuring of properties are of interest. Pivot mapping, which will be explained later in this document in section 3.1, requires a pivot language used for the schema tags, e.g. English. This facilitates a common understanding among the developers about the intention of properties. Restructuring of properties can bridge problems arising from differing conventions. For example, when searching for a first name, it is more useful to have a property "first name". If the name is to be used for an address label, it might be useful to have a ready-to-use full name as part of the address property.

The translation of entire sections of text is out of the scope of Semantic Interoperability Assets (SIA) in the context of SEMIC.EU. Some texts, however, exist in translation, e.g., in the Internet. This means the data could be semantically tagged by references to these translations. For example: "*EUR-Lex provides direct free access to European Union law, making it possible to consult the Official Journal of the European Union as well as the treaties, legislation, case law, and legislative proposals*"[1]. All texts are available in the official languages of the European Union and permit referencing of the local translation. Another example is the DGT Multilingual Translation Memory of the Acquis Communautaire (DGT-TM): "*A translation memory is a collection of small text segments and their translation. These segments can be sentences or sentence parts*".[2]

2.2. Controlled Vocabulary

A lot of data includes properties that have a restricted value domain where all values can be enumerated, such as a closed set of country names. This type of properties can use code lists to represent the values. Typically, the users of the final applications can only select specific and predefined values from a list. From the multilingual point of view, two aspects are of special interest:

- For the same value of the property, the user sees localized labels, depending on the language selected for the user interface.
- If there are two localised schemata, there should be a predefined mapping of values related to such properties. The property values must be mapped while preserving the meaning of the terms used.

Both cases are supported by multilingual thesauri [4][5]. A thesaurus is a list of relevant concepts of an application domain. A concept can be a single-word term or a multi-word compound term. The concepts can be hierarchically organised using the standard relationships "Broader Term" (BT) for generalisation or "Narrower Term" (NT) for specialisation. Further, the predefined relationship "Related Term" (RT) can be used for synonyms. A multilingual thesaurus includes a descriptor of a concept for each language [6]. Usually, one language will become the dominant language, which includes all concepts. Currently, technical thesauri are replaced by ontologies, which represent a more general approach. There are still multilingual thesauri for several application domains available that

[1] http://eur-lex.europa.eu

[2] http://langtech.jrc.it/DGT-TM.html

can be used, e.g., "*Eurovoc is a multilingual thesaurus covering the fields in which the European Communities are active; it provides a means of indexing the documents in the documentation systems of the European institutions and of their users. The European Parliament, the Office for Official Publications of the European Communities, the national and regional parliaments in Europe, some national government departments and European organisations are currently using this controlled vocabulary*".[3]

Controlled vocabularies can only handle the translation of single terms. The translation of longer texts is out of the scope of semantic interoperability assets in the context of SEMIC.EU, as mentioned above. However, semantic tagging using keywords from controlled vocabularies is of interest. Text categorisation using algorithms from machine learning allows identifying predefined topics from a text that could be used to give the user a general idea of the content of the analysed text [7].

2.3. European Union Commissioner for Multilingualism

In 2007, the EU created the new position of an EU Commissioner for Multilingualism. "*The European Union is founded on 'unity in diversity': diversity of cultures, customs, beliefs and languages. Linguistic diversity is a particularly valuable feature of the European Union*".[4] On this website, Leonard Orban, the current EU Commissioner for Multilingualism, has published the following objectives:

- To "*... encourage language learning and promoting linguistic diversity in society*;
- *to promote a healthy multilingual economy, and*
- *to give citizens access to European Union legislation in their own languages*".

The following statement underpins the Commissioner's objectives: "*The European Commission needs to deliver results for citizens, and we need to communicate with you in a language you can understand. Promoting multilingualism is an excellent way to bring European citizens closer to each other. To give you access to information and to contributing your views.*"

Although the major projects of the Commission focus on "learning languages" and "promoting multilingualism", the impacts for the objectives of SEMIC.EU are obvious. For various reasons, pan-European data exchange between various partners in Europe definitely cannot be realised via a strict standardisation of data structures and message formats. In fact, it must support the multilingual mapping and translation of content between the native formats. The citizens of Europe should have access to pan-European data in their own languages. This objective is underpinned by the statement in EIF 2.0 [8]: "*… pan-European eGovernment services which are intended for all European citizens or businesses users must be made available to them in all of the official EU languages.*"

[3] http://europa.eu/eurovoc/

[4] http://ec.europa.eu/commission_barroso/orban/policies/policies_en.htm

3. THE CONCEPTUAL BASIS FOR PAN-EUROPEAN INTEROPERABILITY

3.1. Fundamentals of Pivot structuring

The SEMIC.EU project supports its members, all of whom are European organisations, in creating specifications for information structures that can be shared and understood by every member of the community. These specifications will serve as a base for developing applications that mainly deal with the exchange of information items concerning organisations and citizens on the pan-European level.

At the national level, the comprehension of SEMIC.EU specifications is a prerequisite for the ability to transform the information items into well-defined national information structures. From the more general point of view, transformations of information items are analogous to the problem of translation among multiple languages. Whereas in bilingual environments a bidirectional translation has to be performed, multilingual environments demand techniques that are more sophisticated in order to avoid the costs of multiple bidirectional translations.

In order to solve this problem, SEMIC.EU relies on a well-known principle within the area of computational linguistics and machine translation, namely the introduction of "*interlinguas*", also known as pivot languages. A pivot language serves as an intermediary language and can be artificial or natural. Examples of an artificial pivot language are Esperanto [9] or Interlingua [10].

3.1.1. Basic Structure

Assuming four (work) languages, six translation pairs are needed if the pivot approach is not employed. The idea is to avoid a pair-wise translation by introducing a pivot language, which serves as an intermediary language. Figure 3 shows the principle of a pivot language. To translate from language A to language B, A is translated into the pivot language and the result is subsequently translated to language B. It is clear that one needs to define only four translation pairs, and in general the effort is only N compared to N*(N-1)/2 in the non-pivot case, assuming N work languages. This would lead to a combinatorial explosion if many languages were involved. If more than three work languages are involved, the number of translations is smaller in the pivot case.

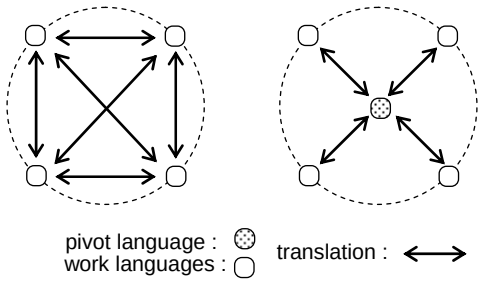

Figure 3: Pivot Structuring

A distinction has to be made between the construction and application of artificial pivot languages on the one hand and natural languages on the other. In the context of semantic interoperability, the focus is set on artificial pivot languages because they are more suited for translating information structures. Both natural pivot languages and artificial ones can serve

as the base for translating natural-language portions in the form of textual documentations and textual requirement definitions.

3.1.2. Applications

An example of the application of an artificial language as a pivot language is the language Interlingua. Interlingua was created by Alexander Gode in 1952. The bases of Interlingua are the Romanic languages and English. Interlingua is also used as a congress language within a small community. For example, several medical congresses, among them the Second World Cardiological Congress in Washington, D.C., 1954, used Interlingua for written summaries. From this viewpoint, Interlingua can be seen as a modern version of Latin, which was used as a lingua franca by Europe's educated class after the demise of the Western Roman Empire.

One of the requirements of an artificial language that serves as an intermediary language is that, ceteris paribus, it should be easy to learn by a huge number of community members. Therefore, it has to have simple grammar and a clear meaning. As previously mentioned, Interlingua has been used as a pivot language at international conferences. The magazine "*Panorama in Interlingua*" proposed it as a pivot language for the European Union in 2006.

Universal Networking Language (UNL [11]) is an artificial language suited for the representation of entities similar to information structures. UNL is an international project of the United Nations University/Institute for Advanced Studies. The aim is to develop a language that encompasses a wide range of domains ranging from academia to tourism and media. It is designed as a formal mathematical language that is able to function effectively while employing only a small set of constructs represented in nearly all languages of the world. Therefore, it is clear that culture-specific properties of a spoken language are not within the scope of UNL.

UNL has three basic constructors, namely:

- labelled links (binary relations),
- universal words, and
- attributes.

Sentences are built by stating labelled links between universal words. The set of universal words is restricted to words that are available in every language of the world. The set of labelled links is restricted to predefined relation names in order to avoid an uncontrolled extension/evolution that would destroy the immediate availability of translation services.

The construction of UNL is highly influenced by the knowledge representation community but with the focus on natural languages. A formal knowledge representation language is particularly well suited to serve as a pivot language for highly structured information such as syntactical and conceptual information models. A prominent example is the language KIF (Knowledge Interchange Format), which can be used to represent information models in the context of interoperability [12]. KIF also has the ability to represent ontology languages like the Resource Description Framework (RDF) [13]. Both RDF and KIF are based on a formal calculus with first-order predicate-logic semantics.

It has to be noted that RDF bears strong resemblance to UNL. In RDF, sentences are represented by statements of the form subject-predicate-object, which is one of the basic forms of sentence structure in natural languages. Analogous to UNL, a predicate is interpreted with a binary relation. In RDF, subjects, predicates, and objects (which are also

called resources) have the form of URLs. In the context of RDF, URLs denote references to real-world objects as well as references to concepts and literals.

For many tasks related to semantic interoperability, it would be overkill to base them on formal calculi like predicate logic. Hence, more-or-less simple representation languages without formal semantics (or at least without the necessity to use formal semantics) can be used as pivot languages. The exchange language XML in conjunction with its accompanying schema language XML Schema (XSD) represents one outstanding example. For example, the exchange between different databases is usually done by means of XML files that are exported from one database and imported into another.

3.2. Issues on Semantics

3.2.1. The Meaning of Semantics

There is a long tradition of studying the nature and form of grammars and languages for the purpose of representing them in a way that can be "understood" by machines. As long as language use and translation only involves human actors, the definition and representation of language semantics is not of primary interest because human actors are inherently able to avoid semantic mismatches. However, in the context of applying information technology for information exchange and using IT systems for collaboration and mutual comprehension, both the nature and representation of semantics have to be considered.

In the area of information technology, there is a long tradition of formalising semantics. The semantics of computer languages is usually described by denotational semantics, which formalises input/output relations of a program, or operational semantics, which formalises state transitions corresponding to each construct. Due to the fact that information models define the structure and content of information instead of describing behaviour or actions, semantics of information models are mostly expressed by set theoretic interpretations.

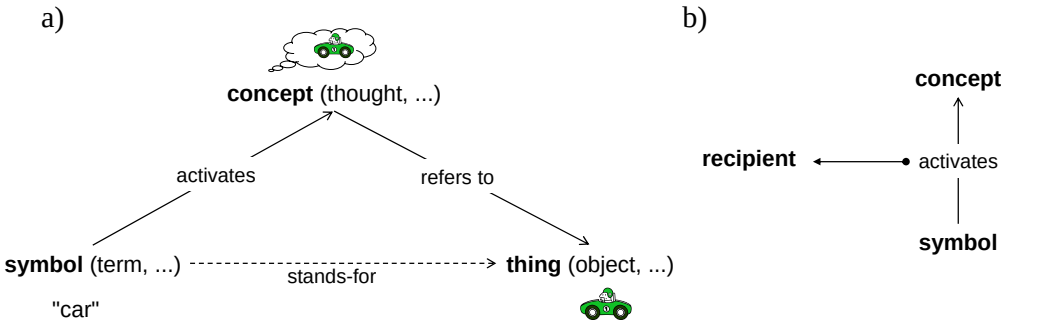

Figure 4: Semiotic Triangle

In the context of SEMIC.EU, the focus is set on the semantics of information models instead of behavioural semantics of computer programs. Information models (also called conceptual models) have a strong relationship to the area of language semantics. The relationship between them is given by the well-known semiotic triangle of Ogden and Richards [14]. The elements of the semiotic triangle are:

- symbol,
- concept, and
- thing.

Figure 4 shows the arrangement of these elements. A symbol is a syntactical entity that activates a concept or thought at the recipient of this symbol - the meaning of the symbol. The concept refers to the real-world thing or object. More specifically, two ISO standards define the triangles entities as follows:

A concept is "*a unit of thought constituted through abstraction on the basis of properties common to a set of objects*" [ISO 5963]. A symbol is a "*designation of a defined concept in a special language by a linguistic expression*" [ISO 1087]. An object is "*any part of the perceivable or conceivable world*" [ISO 1087]. Objects can be material or immaterial.

In the context of information models and their instances, the emphasis is placed on the relationship between symbols and concepts. In order to denote the fact that the symbol/concept pair stands in relationship to a recipient, i.e., the symbol activates a concept at a recipient, these three entities can be connected as shown in Figure 4.

3.2.2. Semantics-Preserving Pivot Mappings

It is clear that translations between languages should not change the semantics of the sentences being translated. In the case of information models, these translations are called semantics-preserving mappings. The relationship between semantics and pivot structuring is exemplified in Figure 5 (a), where the corresponding diagrammatical parts of Figure 5 (b) are integrated.

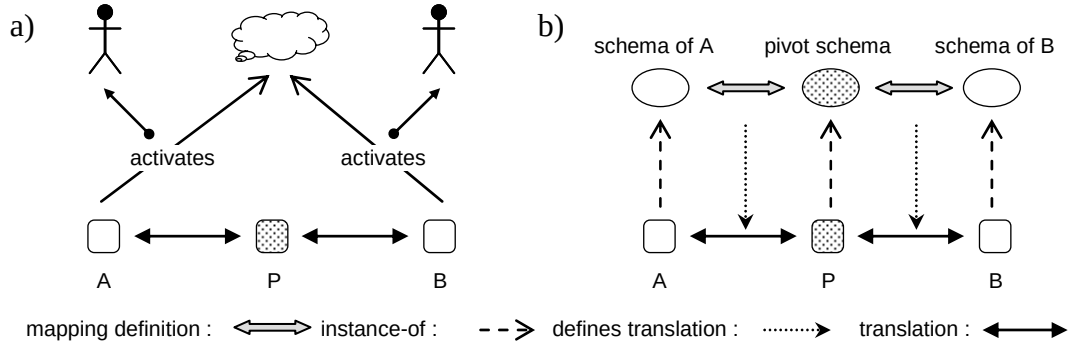

Figure 5: Semantics-Preserving Mappings

The mapping between information item A and B (and vice versa) is done via the pivot structure and has to be semantics-preserving. Semantics preservation is symbolised by the unique concept (cloud), which is activated at both recipients.

A semantics-preserving mapping usually has to be defined over the information models for which the information items have to be valid. Information models in the context of SEMIC.EU can be defined via syntactical modelling languages like XML or conceptual modelling languages like UML, entity-relationship diagrams, etc. In both cases, appropriate mappings, which have to be defined over these models (schemata), specify how the structure and content of corresponding source information items are mapped to target information items. Figure 5 shows the interrelations of all participating entities assuming two recipients. The diagram contains two mapping definitions that map the schema of A and the schema of B to the pivot schema (and vice versa). According to these definitions, translations are performed between corresponding instances.

The two most important entities in Figure 5 are:

- the mapping definitions and

- the pivot schema.

In principal there is no restriction concerning languages that can be used for defining mappings and schemata. In general, five languages could be used or even seven if a different language is chosen for each mapping direction. For instance, a possible scenario could be the usage of the following assignments:

- UML/XMI as the schema language of A,
- RDF Schema (RDFS) as the schema language of P, and
- Data Definition Language (SQL/DDL for relational data bases) as the schema language of B.

Once the schema language is chosen, there are different possibilities to choose the languages for the mappings. Taking into account that one language has to be selected for each direction, there are four mapping parts. For instance, within the scenario sketched above one could choose:

- XQuery [15] for the mapping definition from A to P,
- SPARQL [16] for the mapping definition from P to A,
- TRIPLE [17] for the mapping definition from P to B, and
- SQL for the mapping definition from B to P.

It has to be noted that all mapping languages have to contain constructs that facilitate the construction of structures in the target languages. For example, the mapping definition from B to P assumes constructs within SQL (usually realised by means of SQL/Stored Procedures) to generate RDF graphs.

The above scenario elucidates the general architecture concerning all participating entities presented in Figure 5 (b). It is obvious that a vast diversity of languages could be used in the case of more than two recipients. Furthermore, the above-mentioned scenario acts on the assumption that a direct mapping dependent on the application is preferred, as exemplified by the SQL database from which the information items are directly mapped to RDF structures. However, a more flexible way of defining mappings would be a chained mapping that can be viewed as an internal pivot structuring within the organisation of one recipient.

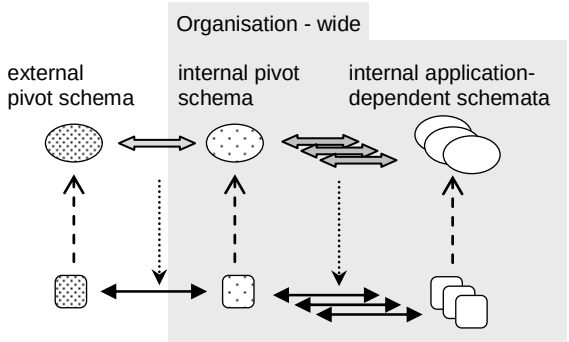

Figure 6: Chained Pivot Architectures

Figure 5 describes the architecture for internal pivot structuring dependent on organisation-wide usage of applications. Chained pivot architectures feature four main advantages:

- improved intra-organisational information integration,
- role separation,

- low coupling and high cohesion, and
- high extensibility.

The first advantage is that the introduction of an internal pivot schema triggers the improvement of information integration within the borders of an organisation. Once an internal pivot schema is defined, several applications can use this schema for the exchange of information. The second advantage of an internal pivot schema is a role separation related to the connection of an organisation to a SEMIC.EU project. In the case of an unchained pivot architecture, all roles responsible for SEMIC.EU-relevant applications have to deal with the definition of mappings between the external pivot schema and the application-dependent schemata. It is obvious that in the case of a chained pivot architecture, only one special role has to be created which is responsible for the mapping between the pivot schema of SEMIC.EU and the pivot schema within the organisation.

Whereas the advantages above mainly concern organisational issues, the remaining advantages are of a more technical nature. The third advantage adopts the well-known principles of low coupling and high cohesion stemming from object-oriented design and development. This is achieved by the single connection between the external and internal pivot schema. That is, the internal pivot schema works as a single interface from the organisation to the external world instead of having as many interfaces as applications. If semantics-changing modifications of the external pivot schema are carried out, only one mapping has to be updated.

The last advantage relies on a similar argumentation. If collaboration with SEMIC.EU leads to an extension of another asset, only one additional mapping has to be defined instead of N mappings, based on the assumption of N relevant applications within the organisation. In this sense, the internal pivot schema acts as the central entity connecting several SEMIC.EU projects with several intra-organisational applications.

Whereas a diversity of languages and systems can be applied within organisations, depending on the needs of the application domains, at the SEMIC.EU level, it seems more appropriate to choose a schema language and a mapping definition language that are widespread and have a high acceptance within the IT-related EU community. One candidate for exchanging information and defining schemata is the XML standard. In this case, XSD can be chosen as the schema language and XQuery or XSLT as a mapping definition language for SEMIC.EU-relevant projects, in order to be aligned with the standards.

Another standard for information models is UML, which is based on the principles of object-oriented design and information modelling according to the entity-relationship model. Although it is primarily a diagrammatic modelling language, it has proper semantics and is well representable by software tools. If a UML information model, which of course acts as a schema in the sense described so far, is chosen as the pivot schema, then the mapping between the pivot schema and the schema of one recipient can be viewed as a transformation from one Platform-Independent Model (PIM) to another. In this case, mapping techniques from the area of model-driven software engineering are candidates.

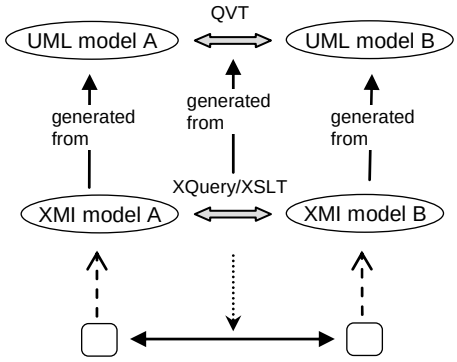

Figure 7: MDA Usage

Model-driven software engineering and the related model-driven architecture (MDA) support the software-development process with languages, techniques, and methods that abstract from concrete platforms and applications [18]. The principles of the MDA approach are based on the same goals as the principles of interoperability: namely, the goals of abstraction from concrete systems and transformability between information items. Two of the main principles are that:

- models described in a well-defined language serve as a basis for understanding conceptual interrelations within enterprise-scale solutions, and
- the construction of systems can be organized around a set of models by imposing a series of transformations between models.

The main entities introduced by the MDA approach are:

- platform-independent models (PIMs) and
- platform-specific models (PSMs).

Based upon these building blocks, several kinds of transformations or mappings can be defined. Including low-level languages such as programming languages, mappings between PIMs, PSMs, and program code can be defined. In the context of the MDA approach, the QVT (query, view, and transformation) framework provides languages to define these mappings [19].

In the context of SEMIC.EU, only the first kind of models is of high relevance. Furthermore, information models rather than models that describe system behaviour, e.g. Statecharts or Petri-nets, fall within the focus of SEMIC.EU. From this viewpoint, PIMs are particularly well suited to serve as pivot schemata. Mapping languages for the mapping from the pivot structure to the information items at the recipient are then given by the QVT approach restricted to PIM-to-PIM mappings.

An essential condition that has to be satisfied when using a mapping language for UML-like schemata, concerns the applicability to instances. More precisely, modelling languages like UML do not provide an explicit instance representation like that provided by modelling languages such as XSD, the Ontology Web Language, RDFS, or SQL/DDL. Therefore, three levels have to be dealt with instead of just the two levels shown in Figure 7. When choosing a tool for defining mappings based on UML-like languages, it has to be ensured that the tool takes care of handling all three levels.

Figure 7 sketches the required connection among the three levels. Here the translation of an instance is defined by two steps. Given two UML models and a mapping defined by QVT, a mapping formulated over the two equivalent XMI schemata is generated from the QVT definition. The second step has to be performed as described in the sections above.

3.2.3. Semantics Preservation and Pivot Vocabularies

So far, we have considered issues related to common linguistic aspects of pivot structuring that are a necessary prerequisite for the preservation of semantics. In order to clarify how structure and content of information items enable semantics preservation, we need to classify information items according to possible base structures. The main elements for constructing information items are:

- literals,
- ranges,
- identifiers, and
- constructs for structuring.

Literals are atomic values also known from programming languages, where they are called simple types. Essentially, they are strings and represent atomic units, i.e., information without any structure that carries meaning. Usually, these values are partitioned by means of simple ranges, which can be ranges for numbers and strings and, for instance, can be represented by the XML schema data types. Furthermore, simple ranges can be introduced dependent of the application domain, e.g., in the context of education, a school type enumerates all possible kinds of schools.

Identifiers are strings that enable the interpretation of values. They can be viewed as a first step towards semantic information attached to values. For example, the value “38” combined with the identifier “age” gives rise to the interpretation that something is thirty-eight years old, assuming that the age of a thing is measured in years. Identifiers also can be used to denote ranges, e.g., the range or type “xsd:date” denotes the set of possible dates in conjunction with XML schema data types.

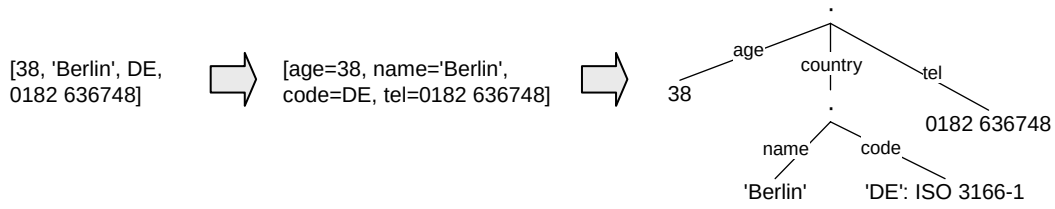

Figure 8: Increasing Semantic Information

According to the foundations of information modelling, the possibility exists to define structures or compositions over literals. These structures also increase the possibility to resolve the semantics of the structured information item. For instance, the composition of the literals “DE” and “Berlin” to the pair (“Berlin”, DE) increases the interpretability of those literals as the well-known city Berlin with its country code. Combining identifiers and structuring leads to information models with rich semantic information. It can easily be seen that all elements for constructing information items introduced so far constitute the core set of elements common to all information-modelling languages.

Figure 8 shows three information items formed with the above-mentioned elements where semantic information increases. First, only raw atomic values are presented where the meaning of the items is highly ambiguous. The second form is known as a property-value list, where identifiers are introduced. The third form - comparable to an XML instance - finally introduces a structure and an identifier for a range (ISO 3166-1), which is an ISO standard for country codes and enables disambiguation of the value ‘DE’.

Based on the abstract structure introduced so far, a mapping now is defined as a transformation from one graph structure to another. In the case of XML, we only have to deal with trees where XQuery or XSLT is applied. Because the mapping is defined between two schemata, mature techniques for schema mapping can be applied. In the context of SEMIC.EU, where the focus is set on semantics preservation, recent schema-mapping techniques combined with ontologies will play an important role.

Semantics preservation is only guaranteed if literals, ranges, structuring, and identifiers activate the same concepts after a mapping is performed (see Figure 8). The mapping-definition process cannot be fully automated yet, due to the fact that machines are not entirely able completely to relate elements of the source and the target schema and instances such that the source and target semantics are the same. However, roles responsible for defining mappings should be supported as much as possible in order to resolve the meaning of information items. This can be done by assigning semantic information to ranges and identifiers as much as possible.

In the context of SEMIC.EU, the pivot schemata are the main entities that have to be enriched with semantic information. As described, a schema consists of identifiers like ‘age’ that identify the role of values and identifiers that identify ranges. The former are usually called properties (or attributes) and the latter are called type names (or class names). These two kinds of identifiers are the most preferable candidates for the conveyance of meaning. In addition, relationships between classes like the subsumption relationship (also called sub-class relationship) can be seen as semantic information as well. However, the subsumption relationship is a type of intrinsic semantic information that is well defined by the set-theoretic foundations of information modelling. Recognizing that properties and types carry semantics, they act as key elements for the understanding of the pivot schema in those roles that are responsible for defining the mapping from the pivot schema to the local schema. In other words, a local organisation has to resolve the correct meaning of the schema key elements in order to define a mapping that preserves the semantics of the pivot schema.

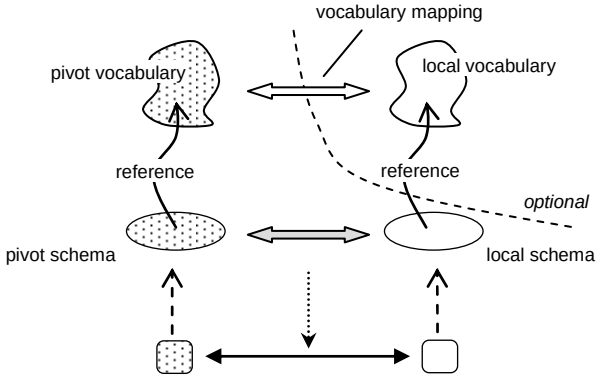

Figure 9: Semantic Enrichment by Pivot Vocabularies

Although schema key elements carry semantics, it would be a benefit to use forms that are more elaborate in order to express semantics, a “controlled vocabulary” being an ideal candidate for this task. The phrase “controlled vocabulary” is a generic term for thesauri, glossaries, ontologies, and taxonomies. Controlled vocabularies help clarify the semantic connection between concepts by defining the relationship between them as well as offering textual descriptions, homonyms, synonyms, and other linguistic correlates. Of course, special languages for defining controlled vocabularies have to be used. For instance in recent years, activities in the area of Semantic Web have yielded ontology languages like OWL [20],

which have enjoyed broad acceptance and are well suited to describe concepts and their relationships.

A pivot schema or a local schema in conjunction with a controlled vocabulary can be seen as an ideal form to enrich schemata with semantic information instead of relying solely on the self-explicability of schema key elements themselves. From a more technical viewpoint, these key elements, namely properties and types, reference elements within the controlled vocabulary. Depending on the degree of precision required and the amount of resources available, a controlled vocabulary can range from simple glossaries to multilingual thesauri or ontologies with formalised rules.

Figure 9 shows a full-fledged application of controlled vocabularies in conjunction with a pivot schema and a local schema. The pivot schema is enriched by semantic information by referencing elements of the pivot vocabulary. It has to be ensured that the pivot vocabulary is understandable by all participating local organisations or recipients. This can be done by creating a multilingual vocabulary or by optionally mapping the pivot vocabulary to a local one. The local organisation resolves the meaning of the key elements of the pivot schema by means of referenced vocabulary elements, from either the pivot vocabulary or the mapped local vocabulary.

The ideal case is that the local schema already references the local vocabulary such that the schema mapping can be derived directly. Should no local vocabulary be available, the meaning is resolved by exploring the multilingual pivot vocabulary. For instance, for the pivot schema elements “country” and “code” (see Figure 8), references to the elements “country code” and “ISO 3166-1” are defined in the pivot vocabulary. These concepts are described multilingual so that they can be understood by all participating local organisations. Once the meaning is resolved, the local role responsible for SEMIC.EU projects is able to define semantics-preserving schema mappings.

4. IMPACTS ON SEMANTIC INTEROPERABILITY ASSETS

4.1. Impacts for Interoperability-related Information Structure

4.1.1. The Pivot Role of Interoperability Assets

One of the primary goals of SEMIC.EU is to set up an organisational and technical framework for multilingual semantic interoperability. The foundations of multilingualism and interoperability and derived concepts described in the preceding sections will be the building blocks for the organisational and technical framework.

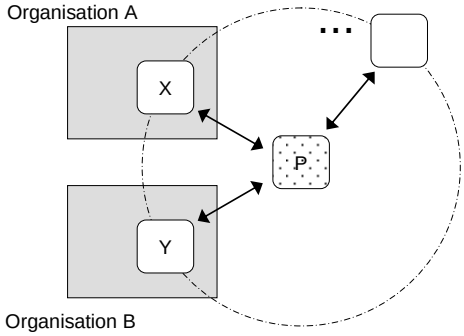

Figure 10: Information Exchange between Organisations

The general requirement in the context of collaboration on a pan-European level for information exchange is semantics preservation by means of pivot structures. Figure 10 shows the pivot architecture extended by organisational borders. Assuming that organisation A and organisation B have a need to exchange information that has to be integrated into existing local applications; the following steps make use of pivot structuring:

- organisation A produces information X,
- X is mapped to the pivot structure P,
- P is mapped to information Y, and
- organisation B uses Y, which has the same semantics as X.

As already mentioned in section 3.1.1, the additional effort required to create and maintain a pivot structure is only beneficial, when more than three organisations participate in the information exchange. The last point assumes that the mapping is semantics-preserving. The question is how SEMIC.EU can facilitate an infrastructure to enable and support semantics-preserving information exchange and ultimately semantic interoperability.

All methods and techniques introduced in the foundations of interoperability should be reflected in the definition of semantic interoperability assets (SIAs). In other words, the question is how the basic structure of SIAs, which only contain artefacts separated in groups, can be extended by elements that support multilingualism and interoperability.

The three main elements of supporting multilingualism and interoperability are:

- pivot vocabularies,
- semantic statements, and
- semantic-preserving mappings.

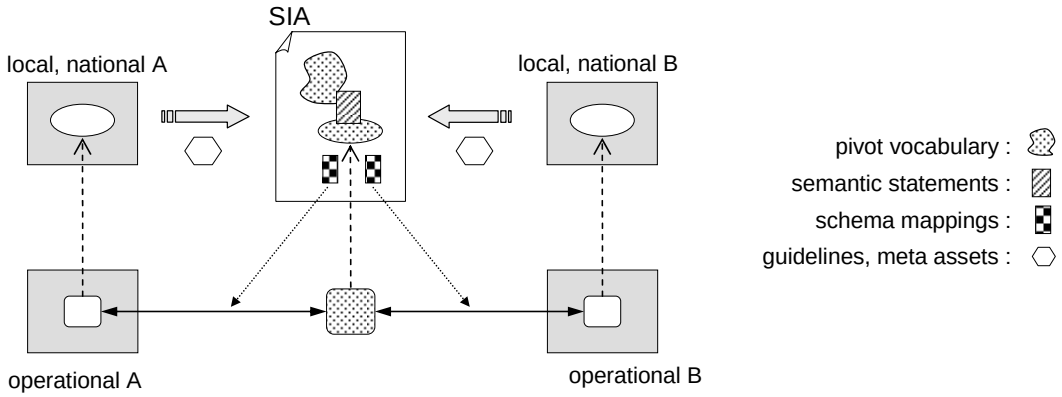

Figure 11: Semantic Interoperability Assets

Pivot vocabularies aid in clarifying the semantics of all concepts related to the artefacts that are needed to understand the whole semantics of the corresponding asset. In the simplest case, these vocabularies are glossaries that explain concepts by means of textual descriptions. In the most elaborate case, these vocabularies would be machine-understandable ontologies that clarify the semantics of concepts by defining special or freely named relations between them. Furthermore, formalised rules target the precise definition of concept relations by logical considerations. The pivot vocabulary has to be multilingual and serves as the base for a SEMIC.EU-wide understanding of asset-related concepts.

The references - from schema elements (identifying names for properties and types) to the pivot vocabulary that disambiguates the meaning of schema elements as much as possible - are formulated by means of so called “semantic statements”. A semantic statement represents the linkage between a schema element and its semantics. In its simplest form, a semantic statement is a multilingual textual description that defines the meaning of a term. Semantics-preserving mappings are schema mappings, i.e., mappings defined over schemata to specify the translation between their instances guaranteeing that the original semantics of the source schema is preserved.

Figure 11 shows the overall architecture for semantic interoperability assets. The main organisational entities are two local organisations, but the principles of introducing and applying SIAs as a pivot structure are of course intended for many participants. Local organisations are split into an operational section and a section responsible for defining and developing local information models needed by applications.

The process of constructing and using SIAs can be outlined as follows: Local administrations participating in a SEMIC.EU asset project construct a common pivot schema on the basis of their local schemata. In this effort, guidelines and meta-assets play an important supporting role. Furthermore, semantic statements are constructed to clarify the meaning of all schema elements by linking them to terms of the pivot vocabulary. Finally, based on these semantic statements, corresponding schema mappings are defined, which leads to a correct translation of related instances.

4.1.2. Artefact Types and Interoperability

The main entity - introduced by the SEMIC.EU project to fulfil the requirements of semantics-preserving information exchange - is called the semantic interoperability asset. The basic structure for assets is a container-like structure that consists of a set of artefacts based on accepted standard formats.

According to the structure introduced in the document "*Vision of the Clearing Process*", the main artefact groups are:

- Requirements,
- Documentation,
- Models, and
- Syntactical Specifications.

Related to artefact and artefact groups are artefact types. In general, one can distinct between:

- types of text-oriented artefacts and
- types of structured artefacts.

For text-oriented artefacts, classical forms of multilinguality form the focus. When dealing with texts, the use of pivot structuring is not the preferred method because the problems raised by semantics-preserving translation of natural language via pivot languages have not yet been solved satisfactorily. However, in the context of semantic interoperability, and especially for structured artefact types belonging to the groups Models and Syntactical Specifications, pivot techniques and methods introduced so far can be applied.

The artefact groups Models and Syntactical Specifications contain schemata as artefacts and correspond to a wide range of artefact types that are used for specifying information and business- process models rooted in the classical entity relationship (ER) and workflow approaches, as well as models for syntactical entities rooted in the SGML approach for defining document grammars. In the context of SEMIC.EU, the following examples for languages and notations used to specify information and business-process models are given:

- Unified Modelling Language (UML),
- Business Process Modelling Notation (BPMN),
- UN/CEFACT Modelling Methodology, and
- Functional Modelling Concept (FMC).

All these languages are primarily diagrammatical languages but usually have an XML representation. Examples of syntactical modelling languages are:

- XML Schema (XSD),
- Relax NG, and
- Schematron.

In contrast to information-model and business-process model languages, they are primarily represented by XML, although diagrammatical forms are used to present a visual model.

These visual models are mainly tool-dependent and may vary among different tool vendors. Two issues are of particular interest when considering interoperability and multilinguality of the itemised languages: Firstly, the concrete schema key elements have to be the subject of investigation. Secondly, the mentioned representation form, which can be a diagrammatical form or given by XML, has to be considered.

The first issue is related to the concrete key elements of the chosen schema language. In general, these key elements are identifiers for ranges or types and identifiers that identify the role of values (properties). In the case of UML and UN/CEFACT, which are UML profiles, key elements are mainly identifiers for:

- classes,
- attributes,
- profiles, and
- roles.

In general, all constructs that are identified by terms are contained in the set of schema key elements. The same is valid for BPMN and FMC. In the context of multilinguality, key elements of BPMN are mainly identifiers or phrases for activities and tasks. FMC comprises key elements like active and passive system components, channels and storages, and classical entity relationship based elements. In the case of XML:

- element tags,
- types, and
- attributes

have to be considered as the main elements for assigning semantic statements.

The second issue, namely the representation form, is important for setting up the linkage between the key elements and the controlled pivot vocabulary by means of semantic statements. In the case of XML as the representation form, semantic statements can be defined by inline references or by using XPath technology to connect schema parts with concepts of the pivot vocabulary. The former way is simpler and can be realized using XML annotations. The more elaborate way is to define an extra structure for determining semantic statements. This structure contains two items similar to association classes in UML. The first item references the XML part of the schema. The second item references the related concept of the pivot vocabulary. At least the first item could be defined using XPath technology whereas the second could be defined by stating a unique concept identifier. In the case of diagrammatical forms, it must be stated that a machine-understandable equivalent representation is available. Depending on the form chosen, special inline annotations or an extra linkage structure has to be defined.

4.2. Impacts on the Development Process

A pan-European Semantic Interoperability Asset will - at least in the majority of all cases - not be created from scratch: The Member States have already gained real-world experience in the field and a significant number of running practical applications exist. The focus of SEMIC.EU is to support the potentially pan-European data exchange between different software implementations of multiple public administrations. To achieve this, common message formats are required that have to be specified in the asset. This means that basic, pan-European, and application-specific data structures need to be specified and fixed. These basic components are used to compose the required messages that will be later exchanged among the application services.

In Figure 12, the process of the initial creation of the asset is shown. It is assumed that a core group of Member States will be established who push on with the Semantic Interoperability Asset and implement a pilot project. It is recommended to start with the current development of one Member State as the seed for the initial draft. It is important to select a pivot language from the start that is used consistently for all artefacts. This concerns identifiers like tag names for the pivot schemata as well as the documentation text. For obvious reasons, English

should probably be selected for pan-European assets, but it can be useful to select another major European language for regional projects. Optionally, individual artefacts, e.g. the documentation, may be translated to the local languages of the Member States if required.

Before the creation of the technical part of the specification, it is important to create the "*OVERVIEW*" and the "*CONCEPT AND DESIGN*" artefact. The Overview describes the purpose of the asset from the business perspective; therefore, it should describe all use cases in detail. As soon as there is consent about the purpose of the asset, it should be registered at SEMIC.EU so the activity becomes visible for other interested parties and potential providers.

The Concept and Design is the technical counterpart that describes the organisational and technical environment in which the asset is intended to be operated later. A very important artefact is the "*REQUIREMENTS*" document, which lists all functions and features (non-functional requirements) to be implemented by the asset. The requirements can be derived from the use cases and typically from the regulations with which they have to comply. The requirements are later referenced as argumentation in the technical artefacts, explaining why something will be specified in this way. Further, the requirements can be checked for consistency to detect contradictions, e.g. deriving from diversity of cultures, legal regulations, and languages.

Consequently, a "*GLOSSARY*" artefact should be created and permanently maintained. For each relevant term from the application domain, a glossary entry should be created that explains the term in general as well as each nation-specific aspect. The terms should include all application-specific terms used in the technical parts, like the properties of the schemata or the controlled vocabulary. The main objective is to achieve a common understanding about the meaning of each term in use. The glossary should be translated to the native language of each Member State. The most interesting parts of the asset are the technical artefacts, like the pivot schemata, pivot mapping, and taxonomies. The pivot schemata are used as an intermediate representation. The mappings specify the translation between the pivot schemata and the local schemata in detail. The validity, consistency, and completeness of the mapping between the schemata and the pivot schemata have to be ascertained. As soon as the initial artefacts are complete, a first draft of the asset should be packaged and published at SEMIC.EU for public review and feedback.

The result of the first milestone will be a first draft of the Semantic Interoperability Asset comprising ideas and technical solutions based on the input of the first Member State but customised for pan-European purposes. The asset is based on the selected pivot language. This means there is an International solution for one Member State that includes the pivot mapping. From it, the other participating Member States may gain a better understanding of what the targeted asset might look like and what additional input is required.

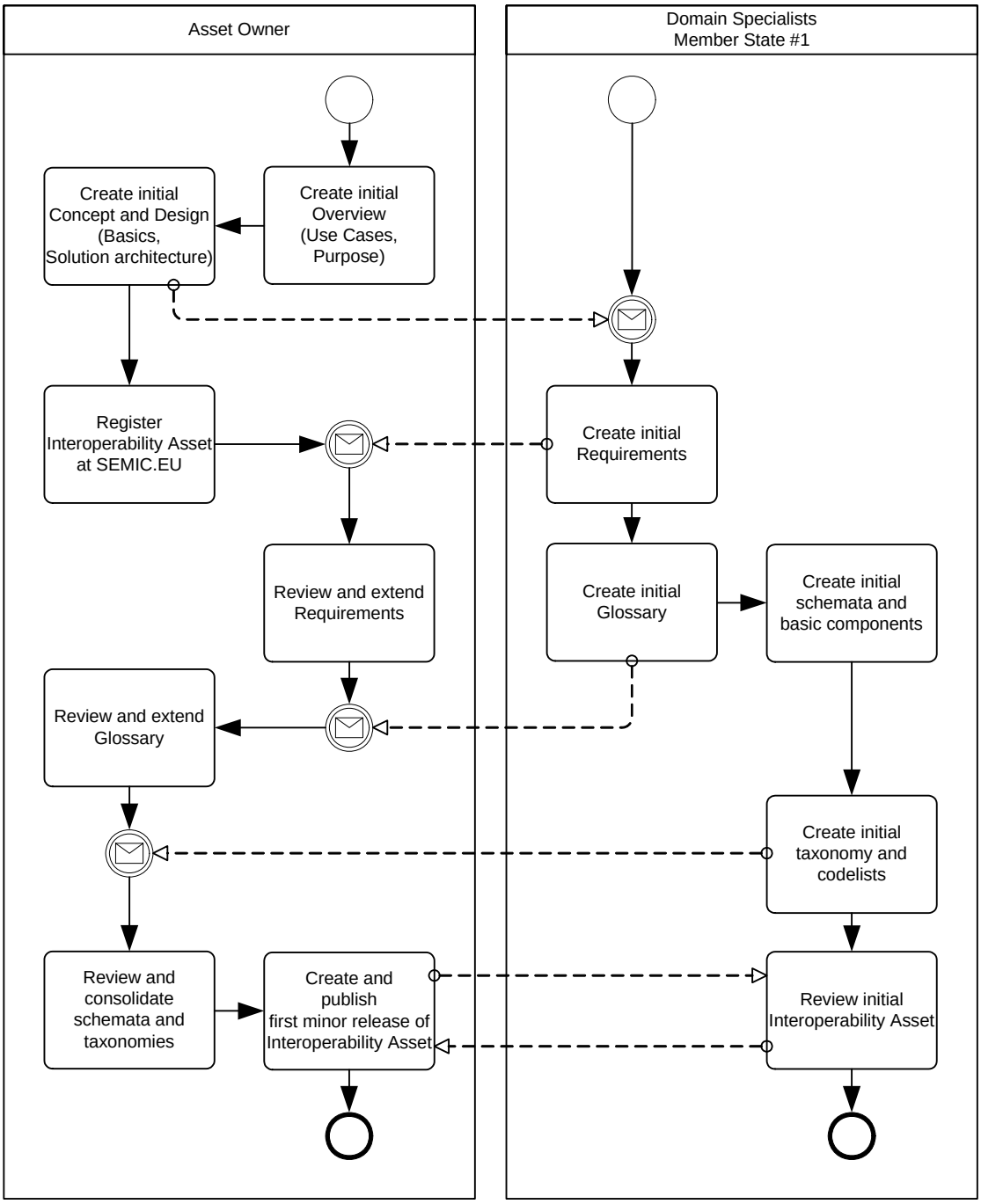

Figure 12: Initial Creation of a Multilingual Interoperability Asset

The next milestone is to create a multilingual solution for multilateral communication inside the core project group. As shown in Figure 12, the Semantic Interoperability Asset will be extended stepwise using the input from the other Member States of the core group. In the first step, a second Member State provides its input and experiences. At this step, the use cases should be approved and gradually improved on-demand to sharpen the business case to be implemented later. The revised asset should enable bilateral communication between the two Member States. At this point, the Member States should provide a mapping between the pivot schemata and/or controlled vocabulary and the available national data structures.

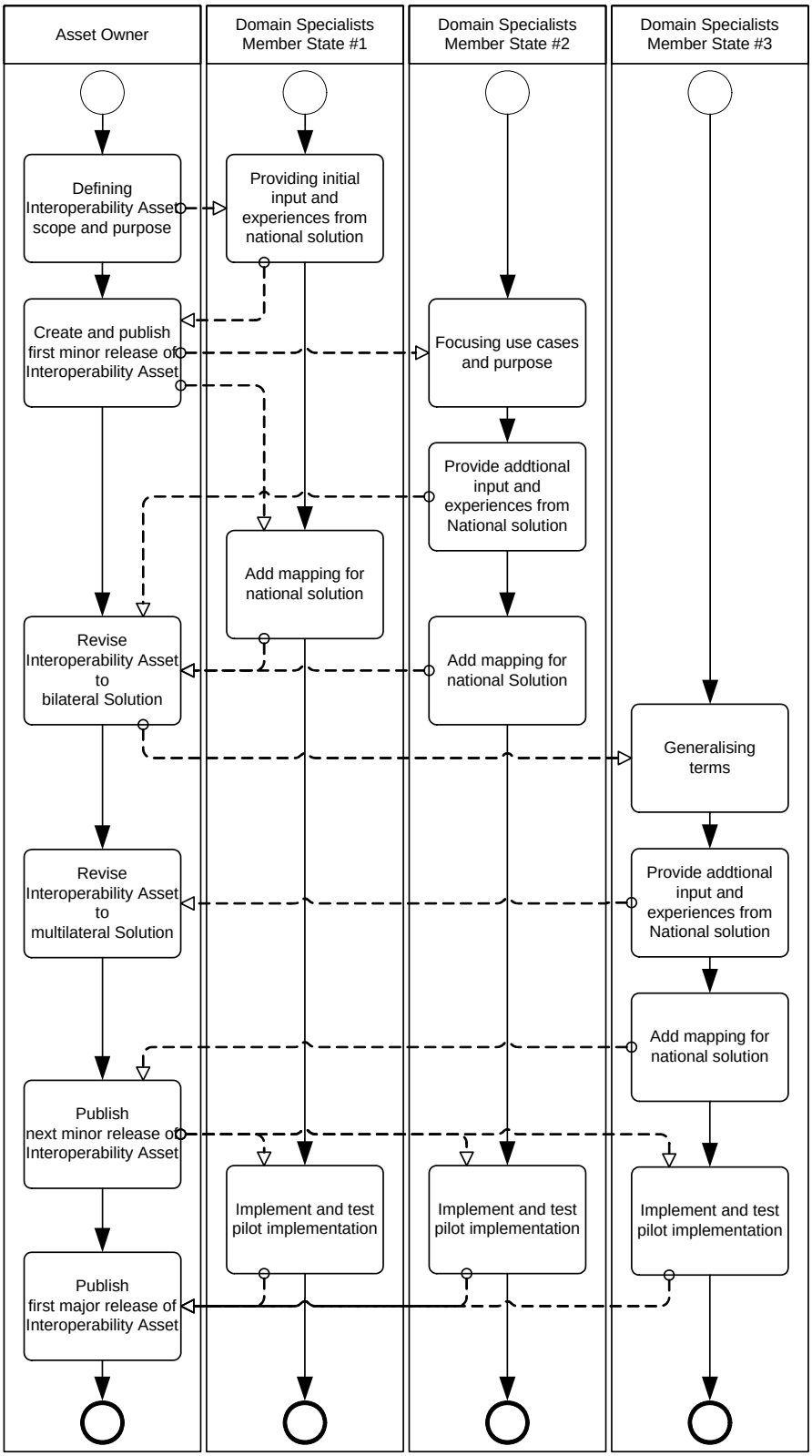

Figure 13: Creation of a Multilingual, Multilateral Interoperability Asset

After the first successful interoperability scenario, a third Member State should be integrated. The focus at this step should be to generalise the solution by approving and potentially improving the terms used. This includes an in-depth review of the glossary, which should contain clear definitions of all application-specific terms. In particular, context-specific

and/or Member-State-specific meanings should be explained unambiguously. The definition of each term should be translated to the languages of the participating Member States. The result should be a multinational approach. After a practical test of the specification to approve the applicability, the Semantic Interoperability Asset should be published as a major release and as a first draft towards a pan-European solution.

As soon as a first stable specification is in place, the rollout of the specification to other Member States should take place. In Figure 14, the overall process to migrate from an ad-hoc data exchange based on national solutions from the Member States to a harmonised data exchange based on a pan-European Semantic Interoperability Asset is shown.

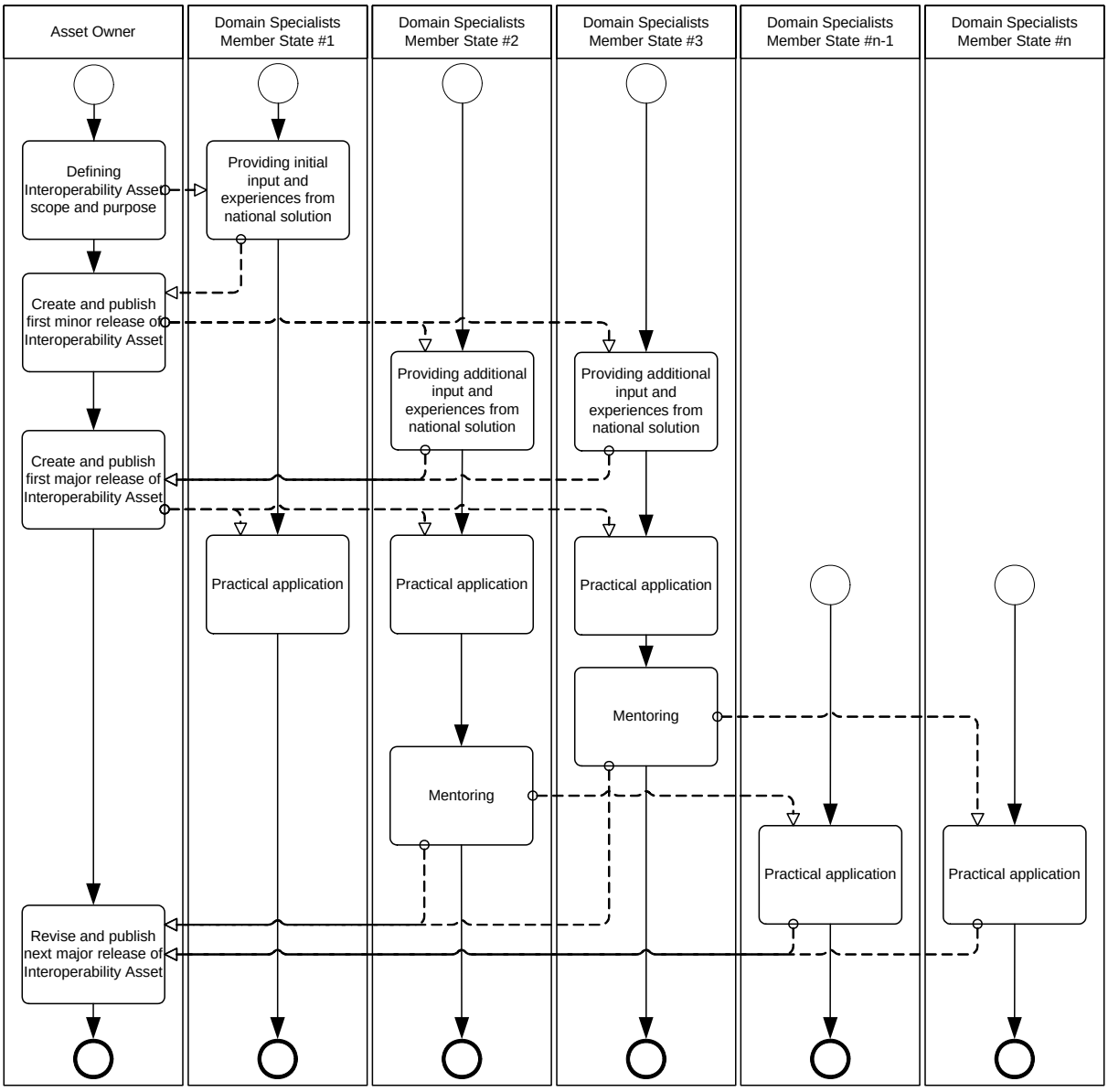

Figure 14: Creation of a Multilingual, pan-European Interoperability Asset

After the first draft of the asset has been published, further input from other interested Member States should be used to improve the asset stepwise. The evolution of the specification will be supported by the maturity model of the SEMIC.EU Clearing Process. From the multilingual point of view, it is important to support more Member States step by step. One suitable method for this is the mentoring model. By making their experience and knowledge readily available, Member States that have already implemented the

interoperability asset in practical applications support Member States that are new to the process. Changes and extensions that may be required will serve as input to improved releases of the asset. The mentoring model should help to roll out the interoperability solution to all partners in a short time frame. It is quite natural that the asset owner, as maintainer of the Semantic Interoperability Asset, will play an ongoing central role in moving the process forward by maintaining the asset in a timely manner, solving conflicts, and providing the required high-level support.

To summarise, the migration process to exchange data among domain-specific applications based on national standards will start with the creation of an initial Semantic Interoperability Asset that supports pivot schemata and/or controlled vocabulary using a pivot language. The next step is to implement a multilateral solution between the core Member States of the project. The complete rollout may utilise the mentoring model, in which experienced Member States already operating the solution are supporting new Member States and add the support for additional native languages, cultures, and legal regulations.

4.3. Impacts of Multilingualism on Asset Users

A stated objective of the EU is that the citizens of Europe be given access to European Union legislation in their own languages and that information from and for the public administration should also be accessible in the various languages of the Member States. Although an increasing number of Europeans are gaining functioning knowledge of the English language, for their quotidian work the corresponding native languages are still ubiquitous and prevalent. This is in particular the case for special domains, which feature and utilise a highly specialised and confined vocabulary, such as justice and public administration. Therefore, it is essential for SEMIC.EU to support multilingualism even for the asset users who are part of projects specifying or implementing Semantic Interoperability Assets.

Providing a multilingual frontend for SEMIC.EU is recommended in order to lower the entrance barriers for new users. The same is true for those portions of the documentation and guidelines that are of interest to all SEMIC.EU users. The decision regarding the languages to be given priority to reduce the initial effort depends mainly on the target groups to be supported at the outset in their role as the “early adaptors” - those who show the greatest interest in working with SEMIC.EU.

However, being able to deal with the SEMIC.EU platform is only one of the various issues for asset users. Much more relevant for them is to find related work - i.e. mainly assets - even if the most interesting assets are provided in the language of the asset owner. As soon as the number of assets managed by the SEMIC.EU platform grows to more than several hundred, simple navigation through the inventory will no longer be practicable. Therefore, it is necessary to support an effective search for assets using keywords, as do typical public search engines. This requires the creation of search indices of the content from the Semantic Interoperability Assets.

One technique is to implement a full-text index from the contents of the artefacts contained in an asset. This is a well-known technique, especially for monolingual content, and it is supported by a variety of tools. In the context of multilingualism, a cross-lingual index is required that makes it possible to find documents although they use a similar term in another language. This would demand that a document be indexed by its own terms and all the translations of these terms. An alternative would be to create an index for each language and

to search for a document in all indices using the translations of the keywords. A precondition for creating a valid index is to know the language of the content to index. This could be solved by providing the language explicitly for each part of the content through the meta-information or using well-known methods from text mining, which can identify the language of a sentence or document using algorithms from machine learning.

Besides the option to index an asset using all the terms contained in its artefacts, it is possible to have a cross-lingual index composed mainly of the most relevant topics of an asset and its artefacts. The topics can be given explicitly by the asset owner through keywords for the asset and for each artefact. The topics can also be automatically extracted from the content of the artefacts by using rule systems and other methods from text mining. An asset has to be indexed using all translations - or even better, multilingual mappings - as provided by the asset itself with its own “controlled vocabulary”, which also contains relevant topics.

Thus, there is clearly a need for an overall concept of indexing Semantic Interoperability Assets. Furthermore, the search strategy and navigation concepts like facetted navigation should be worked out in more detail before being implemented.

Another crucial issue for asset users is the representation of multilingual Semantic Interoperability Assets and their contained artefacts. For example, comments on an XML schema have to be offered in 23 languages - for instance a schema containing translations for each inline comment that might be hard to read. Another option would be to provide 23 schemata containing the comments only in one language for the target group. In this case, it might be difficult to make sure that all translations address the latest valid schema. Another possibility would be to divide the different translations of the schema. The schema itself would only contain the comments using the pivot language and a unique identifier for the comment, and a separate resource file containing the translated comments could be created for each language.

The described solutions can be converted back and forth by tools, but regardless of the method used, providing a common solution for this issue is an absolute necessity to facilitate the work of both the asset developers and the asset users. The solution must also include rules that define how a multilingual asset will be packaged in a common form, e.g. file name conventions and multilingual meta- information.

5. CONCLUSIONS

Leonard Orban, the current EU Commissioner for Multilingualism, defines one of his organisation's objectives clearly: "*Give citizens access to European Union legislation in their own languages.*" Consequently, there is a real demand to provide multilingual support for pan-European solutions. It is not sufficient to develop international solutions and merely using English as the common language.

Semantic interoperability requires preservation of the meaning of the exchanged information in a communication between partners with the concrete goal being the avoidance of ambiguities and/or misunderstandings. The semiotic triangle defines the relationships among concepts (*THOUGHTS*), symbols (*TERMS*), and/or things (*OBJECTS*), explaining that in multilingual contexts, terms from different languages should activate consistent concepts to preserve semantic uniformity. This implies that messages translated from one system to another can use different terms, but should map to the same concept.

On the other side, it should be taken into consideration that a mapping may well be imprecise, incomplete or even ambiguous in some circumstances. In such a case, technical means are required in order to make the user aware of the imprecise translation or incomplete transformation.

The multilingual mapping will usually be included in a semantic interoperability asset to support a pan-European data exchange. The foundation for the multilingual mapping should be a pivot mapping, which implies that in an asset, a pivot schema is provided and there are multilingual mappings between the local schemata of the Member States and the pivot schemata. The multilingual mappings must be validated using the semantic statements defined alongside with the pivot schemata.

It is important to note that for the implementation of pan-European data exchange, local schemata only need to be modified in exceptional cases. The pivot schemata represent a subset of the local schemata adapted to the purpose of the asset and translated to the selected pivot language. As part of the improvement of the SEMIC.EU methodology for the development of semantic interoperability assets, the preferred techniques for the specification of the semantic statements and mappings should be investigated in more detail and recommended. This will help define the concrete specification of the required artefact types for practical use in the assets.

The practical aspect of multilingual mapping of data includes structural changes implemented by techniques as schema mappings and the translation of controlled vocabularies implemented by techniques as code lists, multilingual thesauri, taxonomies, and/or ontologies. An important point here is that some semantic problems can be solved by syntactic methods. Fixed rules that dictate when to use syntactic or semantic methods are not always available. The specific advanced semantic technique to be applied depends on the precision required and the resources available. It is recommended to provide a tutorial on SEMIC.EU with a running example that demonstrates and documents typical cases of multilingual mappings in detail.

A quite special issue is the automatic translation of longer text parts in contrast to the mapping of simple strings or terms. The related techniques of machine translation are the subject of research and not a component of multilingual practice in the context of SEMIC.EU. Instead of seeking to translate texts, it is recommended to support the end user

by using semantic tagging, which adds related topics identified from the text or references to text resources that are similar to the mapped data. The European Commission already provides various rewarding text resources, such as EUR-Lex or the DGT Multilingual Translation Memory. Semantic tagging is an innovative approach and should be a future option for investigation in more detail for practical advisories.

Another important issue is the artefact types that are affected by multilingualism. The detailed specifications of the various artefact types (meta-assets) should explain concisely how to deal with the different aspects and implications of multilingualism. In addition to the special artefact types, e.g. for multilingual mapping, the general cases will be identifiers, single documentation strings, and textual artefact types for the documentation. This also includes more technical issues like how to implement the joint editing of artefacts, e.g. for a documentation string, that include multiple languages in one document. These aspects of the problem should also be considered in the improved SEMIC.EU methodology for Semantic Interoperability Assets.

An example of a systematic development process for multilingual pan-European assets has been developed, outlined, and provided for the interested audience. It demonstrates how a multilingual asset can be created step by step from existing local schemata, starting with an initial Member State's experience and input as a takeoff point and then extending it. Using the mentoring model, the experience and knowledge are transferred inside the project in a distributed manner. Furthermore, it should be checked whether there are other systematic working models to set up a pan-European, multilingual Semantic Interoperability Asset in an effective way that are potentially more applicable under other circumstances.

Ultimately, the SEMIC.EU platform will be significantly affected by multilingualism, and the user interface and part of the documentation need to support all relevant languages. Another important factor is that a multilingual search must be able to find Semantic Interoperability Assets, even if the artefacts are not available in either the user's local language or the pivot language. In the future, this will require providing a cross-lingual full-text index, as well as automatic, multilingual tagging of asset with identified topics of the asset.

To conclude, multilingualism is clearly an issue with a wide range of consequences for SEMIC.EU. As multilingualism is a special topic directly related to the platform's efforts, spreading this know-how and experience to the entire SEMIC.EU community in an effective manner is of utmost importance. In addition to guidelines and white papers that deal with multilingualism, local presentations (road shows) should be organised to talk directly to the target groups of SEMIC.EU at the grassroots level.

The other central question concerns the amount of semantic information that actually needs to be implemented by a Semantic Interoperability Asset. Each project should decide the semantic methods to be used based on its objectives, resources, and know-how. The project needs to consider the degree of scalability necessary for its problem domain's selected semantic method.

Solving the most important issues of multilingualism for pan-European data exchange is a very important and unique selling point for SEMIC.EU. Therefore, this topic should be moved forward rapidly in order to increase the technical advantage of SEMIC.EU and fulfil the expectations of its stakeholders. Fortunately, the study of multilingualism has shown that numerous solutions are already available and can be implemented very soon. Other points still require further investigation to find practical approaches in a timely manner.

Appendix A REFERENCES AND LITERATURE

[1] Held, B., "Infrastructure for pan-European Semantic Interoperability: The IDABC XML Clearinghouse", http://www.egovinterop.net/Res/9/W15%20Held.pdf, 2006.

[2] European Commissioner for Multilingualism, "A New Framework Strategy for Multilingualism", European Commission, http://ec.europa.eu/education/languages/eu-language-policy/doc99_en.htm, 2005.

[3] Park, J. and Ram, S., "Information systems interoperability: What lies beneath?", ACM Transactions on Informartion Systems, 2004

[4] TC 46/SC 9 - Identification and description, "ISO 5964:1985 - Documentation -- Guidelines for the establishment and development of multilingual thesauri", International Standard Organisation, 1985.

[5] TC 46/SC 9 - Identification and description, "ISO 2788:1986 - Documentation -- Guidelines for the establishment and development of monolingual thesauri", International Organization for Standardization, 1986.

[6] Working Group on Guidelines for Multilingual Thesauri, "Guidelines for Multilingual Thesauri", IFLA Classification and Indexing Section, 2005.

[7] Sebastiani, F., "Machine learning in automated text categorization", ACM Computation Surveys, 2002.

[8] European Interoperability Framework v2 (Draft), http://ec.europa.eu/idabc/servlets/Doc?id=31597, November 2008.

[9] Forster, Peter G. “The Esperanto Movement”. The Hague: Mouton Publishers, 1982.

[10] Gode, Alexander, “INTERLINGUA”, The Journal of Communication, Vol. 5, No. 2, 1955.

[11] “The Universal Networking Language (UNL) Specifications”, Version 3.0, UNL centre, UNDL Foundation, 2001.

[12] Genesereth, M. R. and R. E. Fikes, “Knowledge Interchange Format, Version 3.0, Reference Manual”, Logic-92-1. Computer Science Department, Stanford University, 1992.

[13] Lassila, O. and Swick, R. R. (eds), “Resource Description Framework (RDF) Model and Syntax Specification”, W3C, 1999.

[14] Ogden, C. K. and I. A. Richards, “The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism”, 1923.

[15] Boag, S. et al. (eds), “XQuery 1.0: An XML Query Language”, W3C, 2007.

[16] Prud'hommeaux, E. and Seaborne, A. (eds), “SPARQL Query Language for RDF”, W3C, 2008.

[17] Sintek, M. and Decker, S., “Triple - an rdf query, inference, and transformation language”, Deductive Databases and Knowledge Management (DDLP), 2001.

[18] A. Kleppe, J. Warmer, and W. Bast, “MDA Explained. The Model Driven Architecture: Practice and Promise”. Object Technology Series. Addison-Wesley, 2003.

[19] Object Management Group, “MOF 2.0 Query/View/Transformation, v1.0”, OMG, 2008.

[20] McGuinness, L.D. and van Harmelen, F. (eds), “OWL Web Ontology Language - Overview”, W3C, 2004.