Wide Information Network for Risk Management

IST Integrated Project No FP6-511 481

Deliverable D2205.3

Risk management language & communication TOOLS

ONTOLOGY

Instrument: Integrated Project
Thematic Priority: Risk Management
Start date of project: 1st September 2004 **Duration**: 40 months
Project coordinator: Christian Alegre - Thales Alenia Space

Organisation name of lead contractor for this deliverable : MULTH-UMB
Document reference: WIN-UMB-HLI-MULTH-PU-D2205.3, RM language&Communication tools : ONTOLOGY.
Due date of deliverable : T0+36 **Actual submission date** : 07/07/2007
Revision : 2.00

Dissemination level : PU (Public Dissemination)

Author	Company	Date	Signature
BUDIN/Bornemisza	WIN-UMB-HLI-MULTH-VIENNA		BUDIN
Checked by			
BUDIN/GRECIANO	WIN-UMB-HLI-MULTH-SXB		GB/ GG
Approved by			
Christian ALEGRE & Cecile MONFORT	Thales Alenia Space		

IST Integrated Project No FP6-511 481

REVISION RECORD

ISSUE	DATE	UPDATES	AUTHOR
1.00	31/08/2006	Creation and completion for CDR	**BUDIN/**Bornemisza
2.00	07/07/2007	Finalization	**BUDIN/**Bornemisza

ABSTRACT

This document provides an introduction to terminologies as corner stones and pre-requisites for creating ontologies. In a second part the methodological building blocks needed for ontologies in the WIN project are described.

KEYWORDS

Ontologies, terminologies, semantic interoperability

COPYRIGHT

TABLE OF CONTENTS

1 SCOPE, INTRODUCTION, AND OVERVIEW

This document is the deliverable for WP 2200 "Human language interoperability " , task 2205.3 "Ontology"

The scope of this document focuses on the creation of ontologies as part of the work package WP 2200 Human Language Interoperability. It is closely related to other deliverables in the series D2205 1-4 especially to D 2205-3 on frame semantics. It builds on D2201, D2202, D2203, and D2204. It is also related to other work packages in the WIN project.

First, a methodological outline of ontology building is given. On the basis of this generic description more specific requirements as they emerge in the WIN project are specified. The path from terminologies to ontologies is then described in a procedural form. There is a focus on multilingual ontologies.

2 METHODOLOGICAL OUTLINE OF ONTOLOGY BUILDING

Unlike the traditional and age-old field of Ontology (capitalized and singular) as a branch of Philosophy asking questions about the nature of reality, the existence of certain objects in our environment, and which categories we should distinguish for different kinds of objects, ontology engineering is a young off-spring of computer science focusing on producing ontologies (small letter and in plural) that have a primarily digital existence as formal world models designed to represent knowledge about the world in computers and to enable knowledge engineering applications to process this knowledge for various purposes (for historical accounts see for instance Sowa 2000, Gómez-Pérez et al 2003, Budin 2006).

With the Semantic Web coming of age, ontologies are rapidly becoming the structural backbone of Semantic Web applications, providing not only a formal basis of concepts, properties and relations as more static components, but also rules and constraints governing these static components in their dynamic use in concrete situations. This distinction between static and dynamic components of ontologies leads us to the heart of the discussion of what an ontology actually is. Although ontology engineering is very young, we find a lot of divergent definitions representing different views on and approaches to ontology engineering (see Gómez-Pérez et al 2003 for an extensive discussion and systematic comparison of various definitions in the research literature). But all these definitions (especially the frequently quoted definitions by Gruber (1993), Borst (1997), Studer et al (1998), Sowa (2000), and many others) agree on some invariant characteristics of the concept of (an) ontology: a formal, explicit specification of a shared conceptualization, which can be described as an abstract model of relevant concepts that is shared by people and groups in their professional work, which contains explicitly defined concepts and constraints of their use, and which is machine-readable.

Mainstream ontology engineering is using frames and first-order logic to define the basic components of ontologies: (1) classes that represent (abstract or specific) concepts, (2) relations that specify the (many different types of) relations between classes (or associations between concepts), (3) functions specifying arguments in triples in specified relations, (4) axioms expressing constant propositions, and (5) instances representing concrete elements and individual objects.

We also distinguish many different kinds of ontologies, with some lower degree of consensus in the research literature. Again we refer to Gómez-Pérez et al 2003 as the most detailed discussion of existing typologies and the differences between them, arriving as a conclusion to a multi-dimensional typology. The most often mentioned types of ontologies are (1) upper/top-level/foundational ontologies defining the most abstract categories and a categorial system, (2) domain ontologies formalizing (usually existing) terminologies, knowledge organization systems such as thesauri and classification systems, task ontologies specifying certain processes to fulfil specific functions, to perform certain tasks, and application ontologies that are specific to a certain computer application.

Ontologies may show different degrees of formalization and axiomatization. This is crucial for the subsequent procedural description of developing ontologies from terminologies. According to Obrst/Liu (2003), Obrst (2003) and Daconta, M./ Smith, K./ Obrst, L. (2003), we can summarize this discussion as follows:

- "An ontology defines the terms used to describe and represent an area of knowledge (subject matter)
  - An ontology also is the model (set of concepts) for the meaning of those terms
  - An ontology thus defines the vocabulary and the meaning of that vocabulary
- Ontologies are used by people, databases, and applications that need to share domain information
  - Domain: a specific subject area or area of knowledge, like medicine, tool manufacturing, real estate, automobile repair, financial management, etc.
- Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them
  - They encode domain knowledge (modular)
  - Knowledge that spans domains (composable)
  - Make knowledge available (reusable)
- The term ontology has been used to describe models with different degrees of structure (Ontology Spectrum)
  - Less structure**:** Taxonomies (Semio taxonomies, Yahoo hierarchy, biological taxonomy), Database Schemas (many) and metadata schemes (ICML, ebXML, WSDL)
  - More Structure**:** Thesauri (WordNet, CALL, DTIC), Conceptual Models (OO models, UML)
  - Most Structure**:** Logical Theories (Ontolingua, TOVE, CYC, Semantic Web)
- Ontologies are usually expressed in a logic-based language
  - Enabling detailed, sound, meaningful distinctions to be made among the classes, properties, & relations
  - More expressive meaning but maintain "computability"
- Using ontologies, tomorrow's applications can be "intelligent"
  - Work at the human conceptual level
- Ontologies are usually developed using special tools that can model rich semantics" (end of quote)

Maedche (2002) provides a broad procedural methodology in a multi-layered ontology engineering framework with a focus on ontology learning from corpus analysis, ontology merging, ontology refinement, and ontology evaluation.

3 SPECIFIC REQUIREMENTS TO ONTOLOGY BUILDING IN THE WIN PROJECT

For the specific situation of the WIN project, and thus for the purpose of WP 2200 "Human Language Interoperability", the following requirements to ontology building can be formulated:

- WIN is covering a complete risk management cycle and on a multi-risk paradigm. Thus, WIN ontologies must be modeled according to the generic risk management cycle and must be based on a formal model of different types of risk as defined by different domains and application environments. This implies a multi-, if not trans-disciplinary approach to all types of ontologies (not only the domain ontologies where this is most obviously the case). It also implies a dynamic perspective on all types of ontologies to be developed, not only on task ontologies (where again this is most obviously the case). And it implies that all types of ontologies will be needed in WIN (foundational ontologies, domain ontologies, application ontologies, task ontologies) with different degrees of formalization according to the ontology spectrum mentioned above.
- WIN is focusing on architectural, infrastructural and functional aspects for broad-band, real-time, multi-lingual and multi-modal risk and disaster communication. This has many implications for ontology building: WIN ontologies must cover all media, all communication modalities, multiple languages, real-time decision making processes.
- WIN will be an open system and a dynamic network with a lot of co-operation between many different institutions, thus interoperability issues will be crucial for ontology building and use in WIN and in particular for inter-project co-operation (not only with "sister" projects such as ORCHESTRA and OASIS, but also many other risk-related projects and even more projects in other, related fields in environmental information systems. This means that an interoperability framework will be required to position the WIN ontology engineering framework accordingly.

The following procedural model reflects these requirements and translates them into a multi-phase, yet flexible methodology:

4 FROM TERMINOLOGIES TO MULTILINGUAL ONTOLOGIES: A PROCEDURAL DESCRIPTION

Since WIN has to cover the whole ontology spectrum from the lowest to the highest degree of formalization of language resources, the methodology has to include a procedure for computational terminology management as the basis and point of departure. At the same time all types of interoperability have to be taken into account:

4.1 TERMINOLOGY INTERCHANGE AND SEMANTIC INTEROPERABILITY

The following sub-chapters describe in detail the approach to semantic interoperability:

4.2 HISTORICAL DEVELOPMENT OF TERMINOLOGY INTERCHANGE: THE SALT PROJECT

The acronym "SALT" stands for Standards-based Access to multilingual Lexicons and Terminologies. The objective of the SALT project has been to establish a research agenda designed to promote semantic interoperability among terminological information systems (termbases) and to facilitate access to terminological and other linguistic resources. For this purpose a set of tools has been developed to help various user groups, in particular translators, terminology managers, localizers, and technical communicators, but also tools developers, database managers, and language engineers to achieve such goals in their own application environments. The SALT approach has generated a new view on data and database structures, resulting in a modeling framework that allows users to analyze and link terminological resources to each other. In the field of translation and localization technologies, this approach has contributed to the achievement of a basic level of semantic interoperability in heterogeneous information environments.

4.3 MOTIVATION: SEMANTIC INTEROPERABILITY AND ACCESS TO LANGUAGE RESOURCES

Globalization in all spheres of society has led to an increasing need for language technologies, in particular translation and localization technologies, as well as terminology management tools. The availability of and access to lexical resources (general language words and lexical units) and terminological resources (terms and term elements) in the form of dictionaries, terminologies, thesauri, text corpora, and the like is essential for the efficient implementation of any type of language technology. The current language industry is witnessing the convergence of a wide variety of technologies in machine translation, computer-assisted translation, localization, multilingual authoring, and cross-cultural technical communication. These resources are increasingly configured using XML-based information architectures that require a robust, flexible, and widely applicable coding and modeling strategy for designing lexico-terminological language resources and for providing the basis for interoperable information systems. The SALT project has produced just such an XML-format called TBX

One of the key results of this research agenda with respect to the interoperability of multilingual resources is the clear separation of structural modeling (the syntactical level), data category specifications (the semantic level), and constraints (on the pragmatic level) that govern both the data categories and the syntactical structures used in language resources. This theoretical and methodological distinction is exemplified in Figure 1 and operationalized in:

- A coordinated metamodeling and meta-metamodeling method, and
- A toolkit for 1) specifying the semantics of data categories (the meanings of fields in databases, along with pre-specified content values), 2) specifying the constraints that govern the syntactical models and the metadata semantics, and 3) selecting the data categories relevant to a given termbase management environment.

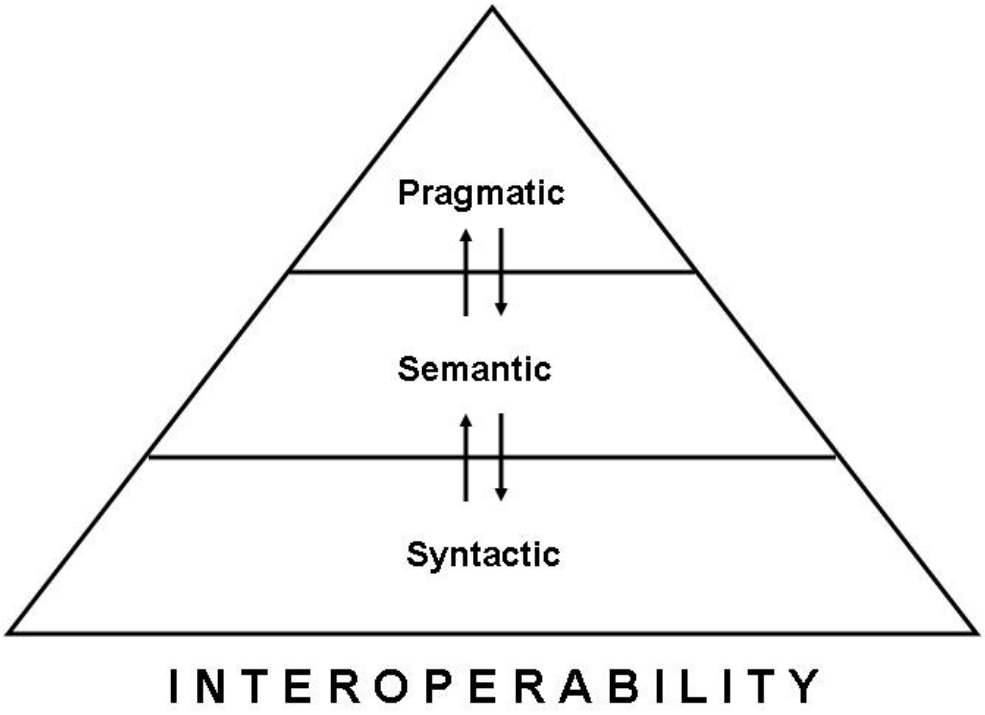


Figure 1: Three levels of interoperability and their interdependencies

The three levels of interoperability shown in the figure ideally match with the three core components of the XLT research agenda: the syntactic high-level structuring framework (TMF), the data category specification mechanism for controlling the semantics in the form of a metadata registry (the new version of ISO 12620), and the pragmatic level of constraints that govern the two other levels and that produces TBX or other terminology markup languages (TMLs). Later we will see how the interoperability required in the development of ontologies for access to a Semantic Web closely mirrors the model for interoperability as follows:

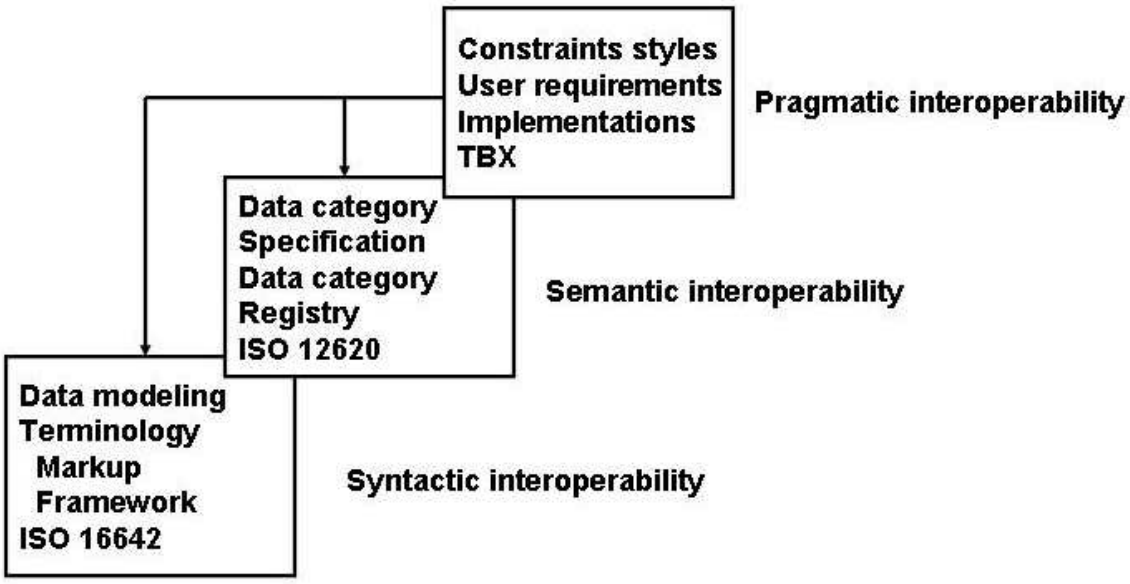


Figure 2: Interoperability and the three levels of the XLT research agenda

4.4 METHODOLOGIES

The work of the project proceeded based on four primary methodologies:

- The analysis of existing data structures in major terminology databases with a goal of mapping application-specific data categories to the data category specifications from ISO 12620
- The derivation of an XML-based specification (TBX) for tools designed to allow different user groups to convert their data structures to and from the generic TBX format.
- The investigation of user requirements (user modeling) in different user communities
- The creation of a "demonstrator"—a set of demonstration tools based on user modeling and mapping activities, along with the implementation of significant components for initial testing.

The aim of ***data analysis*** was to describe in detail existing lex/term data formats and the structures of concrete sample data provided by third parties. The results were applied to the design of general data models, which can be used for processing different data formats. Mapping procedures aim at developing a conceptual mapping scheme for data elements as they are employed in heterogeneous lex/term resources. This process also includes the mapping of heterogeneous ontologies in order to provide a structural basis for data mapping. About 25 different database structures have been analyzed. In the process, the

basic principles inherent in the ISO 11179 family of standards for metadata registries were applied to the structure of ISO 12620 with the goal of turning a mere listing of data categories and their specifications into a true metadata registry. In an iterative process, the initial TBX specifications have been improved and extended according to the results of the data analysis and the user modeling processes.

Multilingual terminological resources (MTRs) are key components of machine translation systems, technical authoring and controlled language systems, translation memories and text alignment systems, corpus linguistics applications, etc. Multilingual terminological resources (Budin/Melby 2000) have become an important type of linguistic resource in addition to speech resources, written resources (full text corpora, lexical corpora) and multimodal (hybrid) resources. The computational paradigm of language management has led to the creation of domain-specific ontologies (Sowa 2000) that organize domain knowledge in large scale information systems. Ontologies categorize data and information, and terminologies contain all the data "ingredients" needed for creating such ontologies while functioning as knowledge organization systems. Hence, MTRs have proven to be very useful for building such domain ontologies.

4.5 THE SALT ARCHITECTURE OF TERMINOLOGY MODELING

These principles for meta-standards and knowledge sharing based on open standards underlie the SALT project. In analogy to the sprawling metadata initiatives such as the Dublin Core (DCMI, part of the World Wide Web Consortium's Resource Description Framework [RDF] standard), and ISO/IEC 11179, a metamodel-based family of formats is now being defined within the SALT project. The SALT approach allows the mapping of many of the existing formats, categorizations, models, ontologies, etc. mentioned above to each other and the transformation of one specific MTR representation into another specific one.

The SALT family of data formats has the following properties:

- It is based on XML, thereby allowing the use of XSL and other XML tools.
- It is modular in its structure, i.e., those parts of an ontology or elements of lexico-terminological information that are actually relevant for a specific target application can be selected and processed using transformation tools.
- A freeware toolkit is available on the SALT server to be downloaded and to be used in projects.
- It is internationalized, i.e., fully Unicode enabled.
- It is end-user oriented and recognizes different user groups of equal importance, e.g., industrial tools developers, service providers, translators, technical writers, localizers, and other "real" end-users.

Any data model and, consequently, any XML representation format for a data model must include three logical components: 1) a set of the data element types that are allowed in the model (as listed in the DCS, the data constraint specification), 2) the permissible content of each data element type, which may be a data type (for example, ISO date) or a list of permissible values for each closed data element type, and 3) the structural relationships that are allowed among the data element instances.

A basic assumption of the SALT project is that no single data model can possibly serve the needs of all groups who access MTRs. Nor would any one format make use of all the data categories in ISO 12620, which is intended to be an exhaustive inventory. When a representation format is processed, e.g., when an exchange file is imported into an application, each data category allowed in the format must be accounted for, including its permissible content. Therefore, user groups are inclined to disallow unneeded data categories from their data models in order to ensure efficiency.

One way to accommodate the needs of various user groups would be to define one complex all-inclusive master format that contains all possible data categories and their values and then to define subsets of that monolithic format. One difficulty with this approach is that such a master format would be necessarily unstable. That is, as each new data category is allowed, the format must change to allow that new data category. The master format must even be modified to allow for one new permissible value for one data category among hundreds. Thus, maintenance of such a format becomes a nightmare. Or, on the other side of the coin, the format could be frozen and not allowed to change, in which case industry will quickly abandon it, for experience has demonstrated that there is a constant need for new categories and new data domain values, even though the current list is *almost* exhaustive.

Another difficulty with the monolithic approach involves writing flexible routines to process an instance of the master format or any subset thereof. Although general-purpose XML parsers can be embedded into end-user applications, error-messages from general-purpose parsers must be contextualized in order to be helpful to a non-expert. This means that the application must understand the XML DTD or schema of the format. The more complex the DTD or schema, the more *expensive* it is for an application to understand it sufficiently well to accommodate a friendly user interface.

The SALT approach separates form and content in a fashion consistent with the international standard for defining terminological formats (ISO 12200). ISO 12200 and ISO 12620 are the form and content components of a family of formats for representing MTRs. ISO 12200 does not define a particular format; instead it defines a family of formats by showing the structural relationships between metadata categories, such as *descriptive element* and *administrative element*, rather than specific data categories, such as *definition*, *contextual example*, or *modification date*. Thus, the structure defined in ISO 12200, even though it must be amended from time to time, is immune to minor changes in data categories and therefore much more stable than the DTD/schema of a monolithic format. Arriving at a content specification for a particular user group may require

considerable advance negotiation, as indicated in the title of ISO 12200. The structure of ISO 12200 combines with a particular negotiated content specification (DCS) to define a particular format.

Current projects within ISO Technical Committee 37 are aimed at

- 1) defining a very high-level metamodel that leaves room for both XML-based representation formats and relational database design, and
- 2) providing for interoperability between specific formats, that is, for bi-directional conversions between formats with little or no loss of information, so long as the content specification is held constant. The object of these approaches is to achieve a so-called "lossless roundtrip," but obviously, if one format makes a distinction between definitions and contextual examples while another format does not, then that distinction will be lost when terminological information is passed through the less nuanced format. No amount of structural manipulation can compensate for incommensurate sets of data categories.

A third activity involves defining specific XML formats based on the MARTIF standard with various specified constraints, coupled with selected subsets of ISO 12620 (DCS), a process that confines itself to the broad possibilities accommodated by the metamodel and that is referred to as *MARTIF with Specified Constraints (MSC)*. Originally introduced as a TC 37/SC 3 activity, this effort reverted to the private sector and has subsequently been incorporated into the TBX standard.

As noted, the SALT project is adopting the ISO approach just described and adding to it elements for representing information from machine translation lexicons and other NLP resources. Furthermore, the SALT project recognizes the need for an approach to designing relational databases that corresponds directly to the metamodel approach to defining XML-based representation formats. Granted, these days XML representations are being used more often as a direct basis for query and processing without passing through a relational database, and thus the distinction between representation formats and processing formats is being blurred. However, this situation simply emphasizes the need for parallel XML and relational database methodologies for MTRs. One such approach to designing relational databases for MTRs, called Reltef™, is freely available to be downloaded (see Reltef) and has been implemented to support central terminological databases in a multinational medical technology company, a university project in Spain, and the United Nations offices in Vienna. The Reltef approach should easily be adapted to object-oriented or hybrid databases.

The integrative picture of the SALT project that emerges from the inclusion of NLP lexicons, relational databases, and a metamodel can be outlined as follows:

- At the highest level, the metamodel level, the abstract structure of MTRs is represented using an application-independent diagramming method such as ORM (Embley, et al. 1992). At this level, the metadata categories are treated as object classes, and the structural aspect of the metamodel shows relationships between object classes. The structure of a data category specification is also given at this level, using some metadata formalism such as RDF, but no particular set of data categories is given except the master inventory in ISO 12620.

- At the intermediate level, the conceptual data-model level, a split occurs reflecting whether the MTR is represented in XML or in a database. Since the emphasis of this paper is the sharing of MTRs, we will discuss the definition of XML data models. All data models are based on the same core structure, which is compatible with the abstract structure in the metamodel. The core structure is expressed as an XML DTD or schema that is compatible with ISO 12200 as amended in a constrained, XML-compliant variant such as TBX. Each data model is defined by the logical combination of the core structure and a particular data constraint specification (DCS). At this intermediate level, a DCS is expressed as an instance of an XML schema that uses tag names that are intuitive to a terminologist while being equivalent to the RDF specification structure defined at the metamodel level.
- At the lowest level, the specific data-model/format level, conceptual data models defined at the intermediate level are instantiated as actual models implemented in database management systems or as XML formats. A single conceptual data model from the intermediate level can have several interoperable formats associated with it at the lowest level. For example, one format may be very similar to the core structure and thus use metadata category-like tag names that are specialized by the value of a *type* attribute while another format may have many more specific tag names and be very similar to the kind of monolithic format described above. One important benefit of the SALT approach is that various subsets of such a monolithic format can be generated automatically by a terminologist who has access to the SALT toolkit but who does not know how to write or modify an XML schema.

For purposes of this discussion, the metamodel level will be identified as Level 1, the conceptual data-model level as Level 2, and the data-model/format level as Level 3. Those familiar with the firstness / secondness / thirdness distinction of the philosopher C.S. Peirce might notice the following analogy (Peirce 1991). Level 1, the metamodel, is associated with firstness in that it represents the potential for many formats but specifies no particular one of them. Level 2, the conceptual data-model level, is the level most closely tied to secondness in that a given DCS, which is the major contribution of level 2, is an expression of the requirements of a particular real-world user. Level 3, the data-model/format level, is connected to thirdness in that a particular data model or format is a set of rules for representation. These rules are abstracted based on the particular user needs that suggested the model in question and can be applied to new situations and sets of data where analogous needs prevail.

The various formats and databases that are implementations of a particular data model are all guaranteed to be interoperable, unlike arbitrary subsets of a monolithic format, and all data models have the same core structure, based on ISO 16642. Thus, even distinct data models are interoperable up to the limits of their ability to map between the data categories and data-category values in their respective data constraint specifications. This interoperability is coupled with diversity to overcome the incongruence that has plagued access to MTRs until now.

5 SEMANTIC INTEROPERABILITY IN A MULTILINGUAL SEMANTIC WEB

Interoperability in translation and localization environments was a major impetus in the original formulation of the SALT mandate and involved a focus on specialized terminologies and NLP lexicons, as discussed above. Ontologies, taxonomies, and thesauri are obvious companion linguistic resources, but their significance has grown exponentially over the course of recent years as their pragmatic role in information and knowledge management in real industrial applications has burgeoned. This trend culminates in efforts within the WWW community to create a more powerful "Semantic Web," where intelligent agents will use ontological information to conduct intelligent searches, assemble relevant information, and even "intuit" concrete conclusions based on logical axioms embedded in online ontologies. By virtue of SALT's capability to "leverage" (i.e., reuse and enhance) critical terminological information for incorporation into the kind of ontologies needed for implementing the Semantic Web, follow-on projects are aiming toward facilitating semantic interoperability in networked digital environments. The following use-case study addresses one such approach involving digital cultural heritage collections.

Given this fundamental triad of syntactic, semantic, and pragmatic types of interoperability among heterogeneous information systems (see Figure 2 above), it is obviously necessary for heuristic and methodological reasons to account for this distinction in practical operations. All three levels depend on each other in their interaction: syntactic interoperability without the other levels would be meaningless and therefore useless. Semantic interoperability without the syntactic component would be amorphous and without structure and linking mechanisms, and it would be too static and context-independent (solipsistic) without the pragmatic type. The pragmatic type without the semantic component would again be meaningless and amorphous without the syntactic element. The classic semiotic triangle of Morris (Morris 1938) fully applies in this context: pragmatics with its ever-shifting, differing views, perspectives, epistemic interests, and application contexts is related to semantics with its different schemes for knowledge organization and different meanings of terms, variant definitions, etc. Both pragmatics and semantics are related to syntax with its different encodings and different forms of representation for the same semantic content.

6 ONTOLOGIES AND MULTILINGUAL TERMINOLOGIES: STRUCTURAL ASPECTS OF SEMANTIC INTEROPERABILITY

Terminological ontologies involve the pragmatic creation of resources such as thesauri and classifications, concept systems, or indexing systems used for knowledge organization, i.e., to order knowledge and retrieve information.

Semantic interoperability involves, among other things:

- General mapping issues (one-to-one, one-to-interchange scheme, etc.)
- Ontology matching approaches
- Evaluation of alignment algorithms
- Granularity issues of representation
- Practical deployment of ontology-related tools
- Cross-linguistic concept matching.

The creation of taxonomies and nomenclatures for ordering knowledge gained during scientific research is based on a long tradition of terminology management that started in the natural sciences. These resources comprise terminologies that have usually been presented in the form of mono- or multilingual specialized vocabularies and dictionaries. For the sake of international communication, these terminologies were standardized on a global level or harmonized when divergent standards already existed. The development of such terminologies has always been an inherent part of the evolution of scientific and technical knowledge – or in other words – terminology management has always been part of knowledge management. In order to share this knowledge in cooperative work spaces, these heterogeneous terminological resources have to be made accessible in a uniform way. So far, the creation of such resources has been governed by local traditions and conventions of lexicographical practice, and the availability of and accessibility to such resources have been limited due to the kinds of legal, commercial, technical, and other reasons cited earlier in this article. The Semantic Web with its new technical developments and standards provides a good foundation for overcoming these barriers and for adapting multilingual terminologies in order to transform them into multilingual ontologies, which in turn will enhance Semantic Web applications such as e-Commerce and e-Learning.

Sowa proposes a useful view on ontologies that are currently proliferating as a result of dynamic research and development initiatives. According to Sowa (Sowa 2000: 492ff) ontology as such is the study of the things (and the categorization of these things) that exist or may exist in a particular domain. Ontology is no longer just a field of philosophy and epistemology, but has been re-discovered as a key perspective for knowledge management. Ontologies (in plural form) are then the result of this study. Sowa distinguishes between informal and formal ontologies. The latter are currently the topic of work in ontology engineering. A terminological ontology is further differentiated by Sowa into prototype-based ontologies and axiomatized ontologies. In 1993 Gruber provided a definition of ontology that has become classic in the field and often cited since then: *An ontology is an explicit specification of a conceptualization* (Gruber 1993). In this definition the term "explicit" obviously means "computational," as the ontologies in question invariably exist as tools for computer-based work (i.e., in the sense of Sowa's formal ontology). Basically we can differentiate between generic, universal ontologies (mostly with a flat structure and mostly focusing on linguistic

aspects) such as WordNet, and domain-specific (mostly deeply structured) ontologies (such as UMLS – the Unified Medical Language System).

But ontologies are to be seen in the wider context of knowledge organization systems: The main purpose of a Knowledge Organization System (KOS) is to support and sustain the constant and ubiquitous human effort of bringing order into chaos in information and communication spaces. In the cultural sector, KOSs are mainly applied for purposes such as

- Electronic Publishing, Web publishing of cultural content
- Communication on cultural content, describing cultural processes, pieces of art, scientific achievements (science is a form of culture!)
- Workflow management, business process engineering, quality management in cultural collection systems
- eContent, i.e., content management (design, generation, storage, updating, dissemination, and publishing of cultural content on the Internet)

Without KOSs, it is impossible to fulfill any of these functions or to achieve any of these goals. The simple reason for this is that content is essentially composed of conceptual structures and their communicative representations in diverse multimedia and multimodal forms, and that the sheer amount of content produced every day has to be organized by conceptual ordering systems that are being created and maintained according to the pragmatic information needs of various user groups.

Two of the most important types of KOSs have developed either

- In the form of thematically structured vocabularies that are increasingly multilingual to meet the linguistic information needs of translators, technical authors, etc., or
- In the form of thesauri and classification systems for library and archival applications that have traditionally been used for indexing (which is another form of ordering) printed materials and increasingly for ordering electronic documents in digital libraries and digital archives, and for retrieving either meta-information on the documents covering certain topics or accessing the pertinent or required documents themselves.

Both types of KOSs are terminologies. The difference between them lies in their different ways of organizing and presenting concept systems and concept representations according to different purposes and user requirements. In the case of classification systems, we have to distinguish between generic, universal systems and domain-specific systems (which thesauri always are), and again most of these systems have become multilingual for reasons of international cooperation and the increasing needs of cross-lingual information retrieval (CLIR).

Looking at the example of communication in the areas of health and medicine, we can observe the full scale of KOSs that are available to different user groups (medical doctors, nurses, and others who deal not only with patients in order to heal them but who also fill in patient records in medical documentation systems where they are unavoidably confronted with classifying such records according to internationally established classification systems (for example ICD, the International Statistical Classification of Diseases and Related Health Problems), or nomenclatures such as SNOMED®, the systematic nomenclature of (human and veterinary) medicine. All these medical

KOSs are based on medical terminology that has developed since classical antiquity and are intrinsically linked to all forms of medical discourse (in research, in hospitals, and in oral and written discourse). These domain-specific KOSs are indispensable for producing medical texts of different types and of different degrees of formalization (e.g., patient records, scientific articles for medical journals, etc.). Medical writing is an example of technical, scientific writing, and medical documentation is a well-established component of medical research and health practice. Cultural knowledge organization systems include thesauri such as the AAT, classification systems such as Iconclass, various domain-specific terminologies (history of art, history of science, musicology, etc.), authority files (place names, artists' names, periods of artistic expressions and styles, etc.) and domain-specific metadata subsets. The domains of environment and risk management include many many knowledge organization systems.

In the context of the ubiquitous use of new information and communication technologies, all forms of technical discourse have been revolutionized. For the last ten years many of the KOSs mentioned above have been turned into ***ontologies*** (notably not in the traditional philosophical sense but in the context of computer systems as formalized ordering systems). The main purpose of such ontologies is to provide a shared understanding of a domain for reasons of communication in the field and of joint action.

In a wider context of the WWW a number of technologies have evolved that enable different user groups with different interests to use the WWW for purposes such as distance education, E-Commerce, computer-supported collaborative work, cyber-science, digital libraries, etc. As noted in the introduction to this section, Tim Berners-Lee summarized this convergence trend in his concept for the Semantic Web. The basic idea is to make content machine-understandable for a number of processing operations in order to be able to provide technologies and tools that enable content providers to enhance the quality of their work and to facilitate the distribution (including marketing) of content specifically adapted to the needs of individual user groups. Initiatives have started to enhance Semantic Web Technologies focusing on *formalizing* (XML, RDF, ontology languages), *grounding* (formalisms and content analysis), *acting* (knowledge discovery, intelligent agents for information filtering, etc.) and *interacting* (visual user interfaces). A number of technologies, methods, and formats have evolved in recent years, mainly based on XML as the common denominator. RDF as a formalized method for representing and visualizing conceptual data structures and ISO 13250, which defines and specifies Topic Maps, also contribute to the array of technology-driven methods for processing and representing knowledge structures in the form of ontologies.

The **SALT** approach of ***sharing multilingual terminological knowledge*** among heterogeneous linguistic data collections is clearly in line with the needs of the WIN project. The conversion of terminology resources and other suitable language resources into fully formalized and axiomatized ontologies is the big challenge in the immediate future in order to provide a solid basis for the Multilingual Semantic Web. Multilinguality invariably involves cultural differences, even when harmonization and standardization efforts have been undertaken in order to eliminate such differences. Requirements with respect to risk management engineering methods based on web technologies involve character set implementation issues (just one keyword here is Unicode) that are crucial for the localization and translation industry. Solving these issues is a prerequisite for real knowledge sharing on a multilingual level. In the language engineering production chain, terminological markup for the creation and value-added enrichment of data structures is a prerequisite for controlled and directed information design, which in turn is the basis for document design and generation, as well as for storing, disseminating, and retrieving information.

We have ready discussed ontologies in terms of the explicit specification of some subject field. Guarino and Giaretta expand this notion to include *re-use across multiple applications* (Guarino and Giaretta, 1995). In the context of most projects, an ontology is a formal and declarative representation which includes the vocabulary (or names) used to refer to the concepts in the chosen subject area and for representing and communicating knowledge about the field in a structured way. It also includes logical statements that describe what the concepts are, and how they are related to each other. Here we have the direct link to the terminological approach to knowledge engineering, which is fundamental to semantic interoperability: Concept-oriented terminologies relate concepts through explicit relationships such as those expressed in ISO 12620 data categories. ISO 12620 allows for different mechanisms for expressing hierarchical arrangements. The difference is that concepts are interrelated not only through their relation to a position in a hierarchical system, but also in any type of terminological concept field. This means that the typology of conceptual relations in a terminology is by far more complex than in thesaurus and classification approaches, which tend to limit themselves to hierarchical relations. The degree of granularity present in terminology is required to maintain semantic richness for content interoperability, especially in culture-dependent domains such as history of culture, arts, language, etc.

As a consequence, the terminological approach advocated here combines all types of knowledge resources, not limited just to thesauri, classification systems and authority files, but extending to multilingual term banks, linguistic resources of any kind, and other knowledge structures. In comparison to the restricted data model used for thesauri, termbases have a much more sophisticated anda variable data modelan be used in this context, as it specifies a framework designed to provide guidance on the basic principles for representing data recorded in any type of terminological data collection

6.1 SEMANTIC INTEROPERABILITY SCENARIOS

Semantic Interoperability can be defined as the ability of information systems to exchange information on the basis of shared, pre-established, and negotiated meanings of terms and expressions. We have already cited various languages that have been developed to accommodate this kind of interchange (KIF, OWL, and various predecessor languages and formats). Semantic interoperability is needed in order to make other types of interoperability work (syntactic, cross-cultural, international, etc.). We can also distinguish different levels of semantic interoperability, on the metadata level (i.e., data about data) in order to facilitate their identification for immediate and unambiguous re-use, retrieval, etc., e.g., Dublin Core and other metadata sets, or we can interact on the object level of data structures themselves.

As far as technical and methodological issues are concerned, the mapping of knowledge organization systems such as thesauri, classification systems, and ontologies is already a big challenge. There are various procedures to distinguish:

- Mapping across different types of resources, of different resources of the same type, and of elements or views of the same resource that are encoded in different ways (maybe with syntactical or expressivity constraints)

- One-to-one mapping, one-to-more-granular mapping, and one-to-less-granular mapping (the most difficult procedure)
- Overlap, federating, merging, and switching

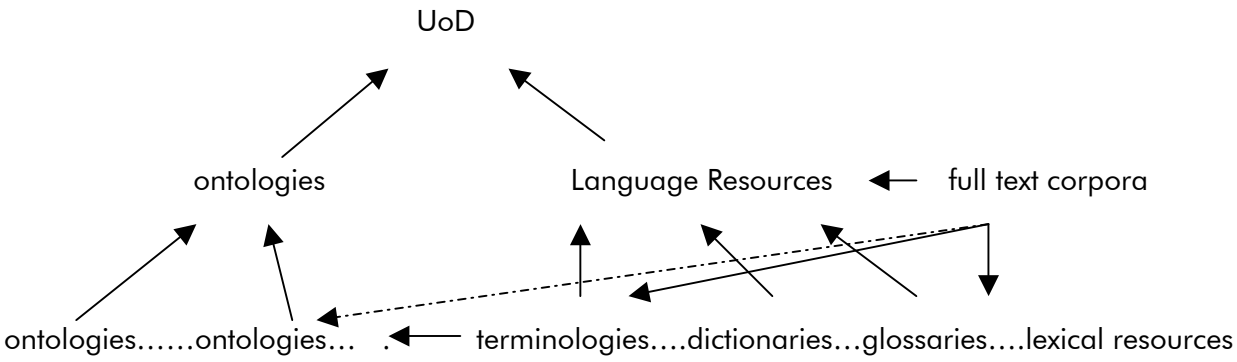
Mapping issues are not trivial, since one-to-one matches are in many cases the exception, not the rule. Different degrees of equivalence are well known in analogy to comparative translation-oriented terminology.

The linguistic approach combined with the traditional controlled vocabulary approach has a number of advantages over controlled vocabularies alone:

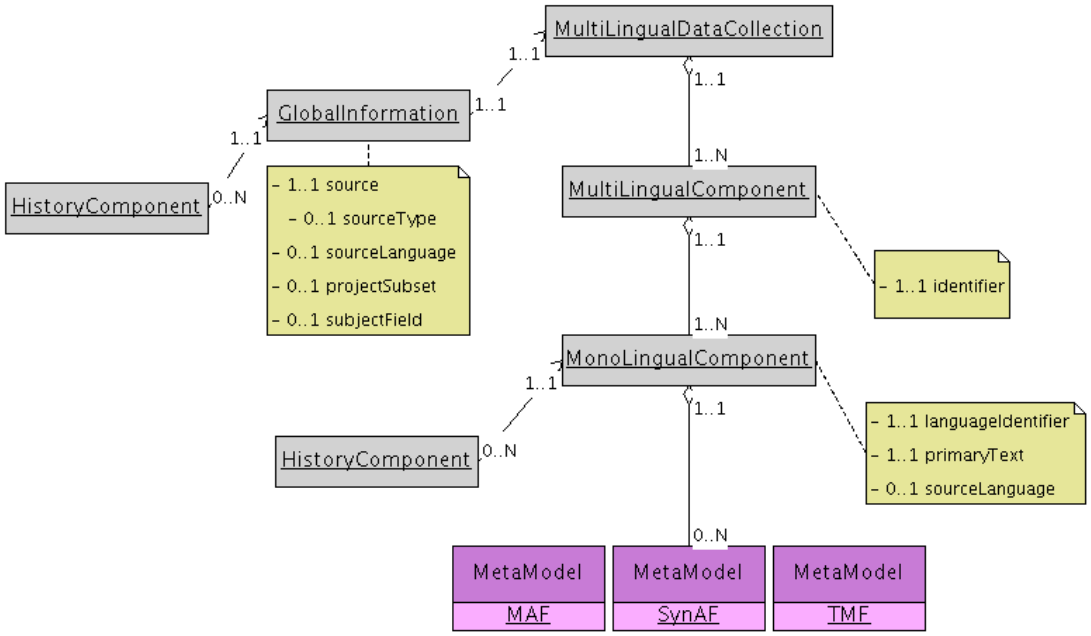
- It captures much more detail fine-grained cultural differences between natural languages that can be very difficult to encode in any knowledge representation or knowledge modeling language.
- It can capture subtle changes of meaning in terms and expressions, particularly in circumstances where the overruling pragmatic reference scheme for constructing meaning suddenly changes.
- In addition, linguistic approaches allow text analysis including automatic term recognition and term collection from texts as well as bilingual text alignments.

The following pages describe some building blocks that are crucial for risk communication systems of WIN.

The following data flows are relevant for the WIN ontology approach: Developing ontologies from other, already existing ontologies, from full-text corpora, from lexical resources such as terminologies and glossaries, all these processes converge in the Universe of Discourse as the overall meta-model and meta-resource.



The Multilingual Information Framework (MLIF) (a draft ISO standard co-authored by the author of this document)integrates a range of standards for managing all types of language resources. It is another building block for the WIN ontology framework for handling multilingual information as relevant to multilingual ontologies:



Ontology mapping is a major issue as well: simple vs complex mapping of multiple ontologies, based on conceptual mapping

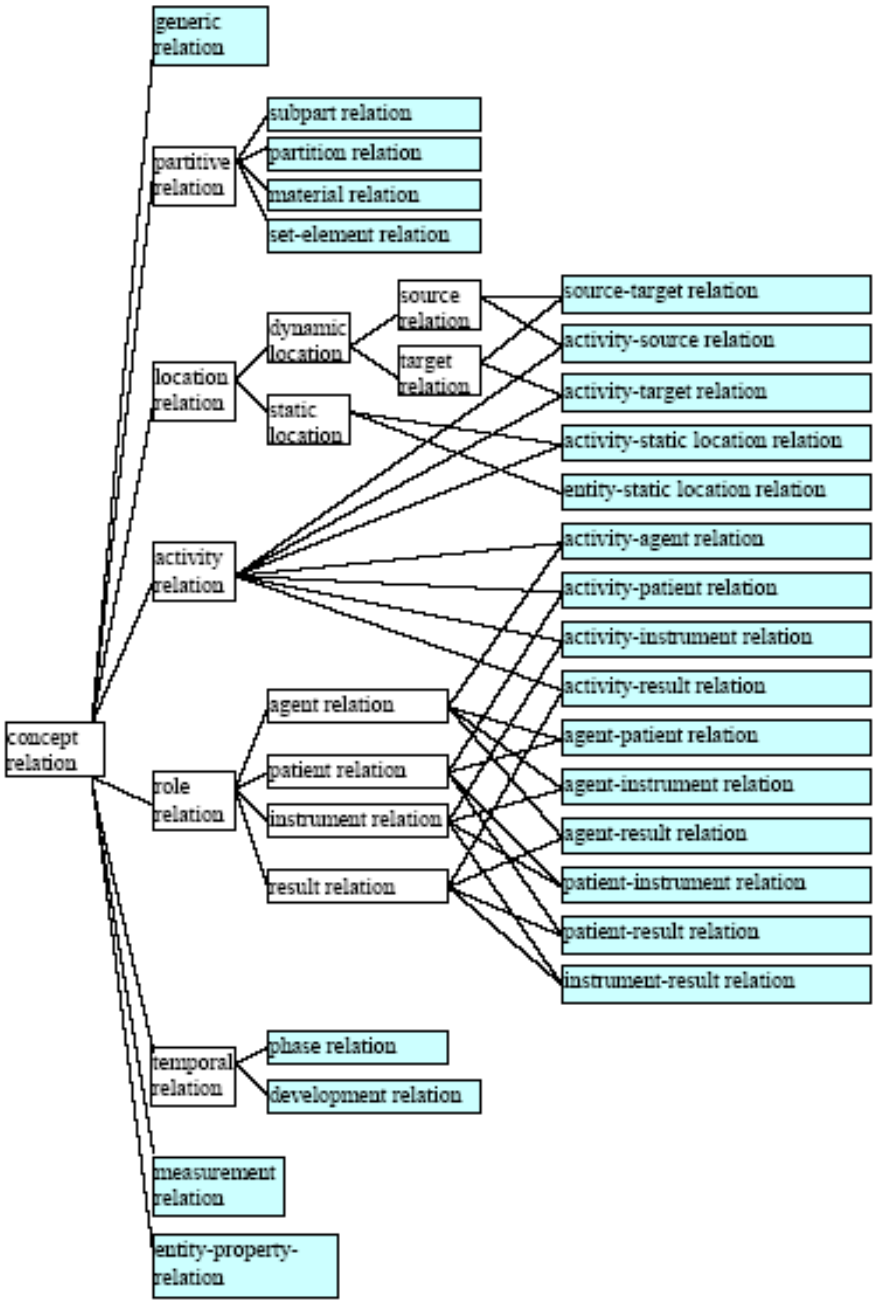
The MOSES project (Pazienza) gives some orientation: Complex mappings vary in the nature of the concepts involved and in the operations that are applied over them, including:

- *restrictions* on classes/associations operated on the range of their attributes
- *aggregations* (on an extensional basis) of multiple classes/associations
- *transformations* between heterogeneous structures of objects (properties vs associations+roles, classes vs instances)
- *join* of associations upon common roles

The question of conceptual relations is a major issue in ontology engineering:

The following is a list of conceptual relations taken from the OntoQuery Concept Relations list: Madsen et al (2001) . It is useful to consider this list for WIN ontology design:

generic relation
entity-property-relation
development relation
phase relation
agent relation
patient relation
instrument relation
result relation
instrument-result relation
patient-result relation
patient-instrument relation
agent-result relation
agent-instrument relation
agent-patient relation
activity-result relation
activity-instrument relation
activity-patient relation
activity-agent relation
role relation
activity relation
static location
entity-static location relation
activity-static location relation
source-target relation
activity-source relation
activity-target relation
target relation
source relation
dynamic location
location relation
concept relation
set-element relation
material relation
subpart relation
partitive relation
partition relation
temporal relation
measurement relation



Due to the fundamental role of Frame Semantics, FrameNet is a crucial database for Ontology Building. Frame semantics is combined with NLP and is then ontologized. (Scheffczyk/Baker/Narayanan 2006)

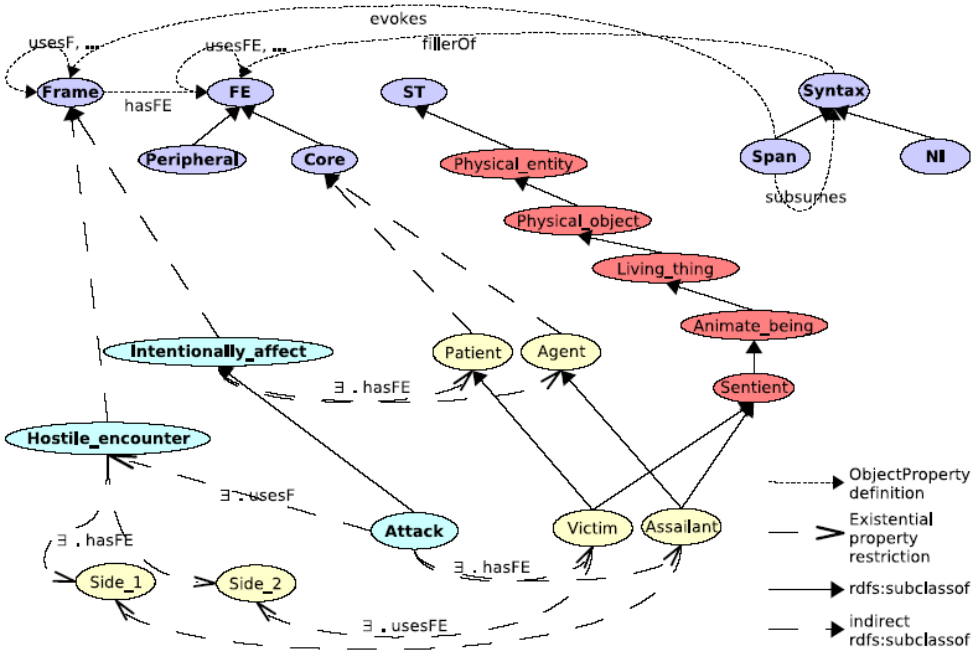


Figure 2: Part of the FrameNet Ontology for the Attack frame and some connected frames.

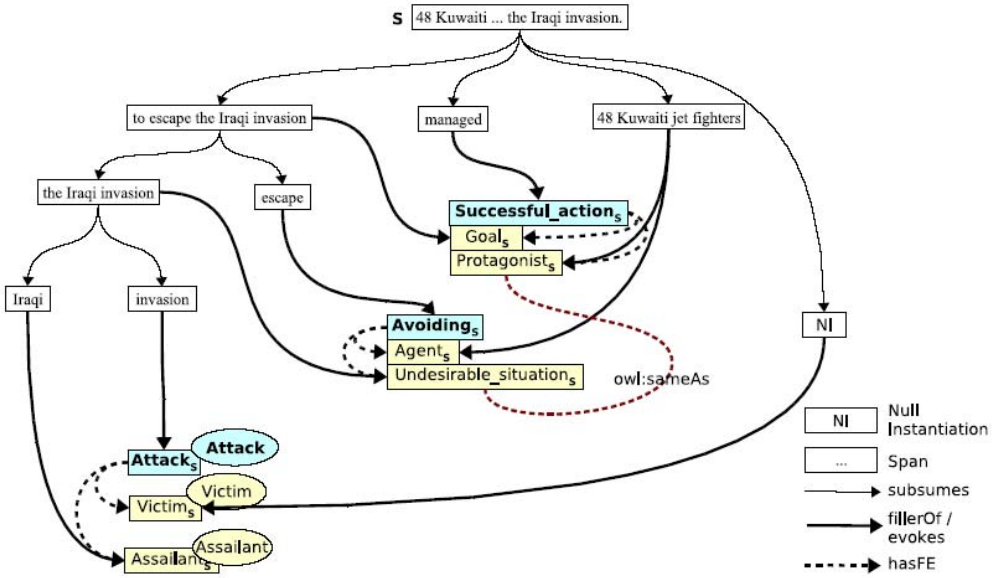


Figure 3: Annotation Ontology for: *48 Kuwaiti jet fighters managed to escape the Iraqi invasion.* (Step 1)

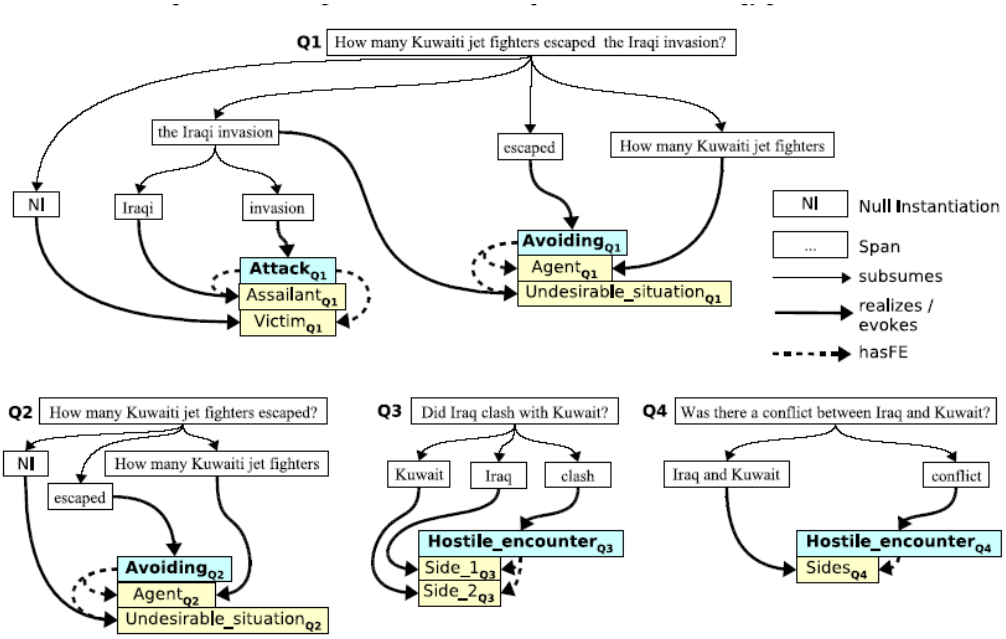


Figure 5: Abridged Annotation Ontologies for example questions

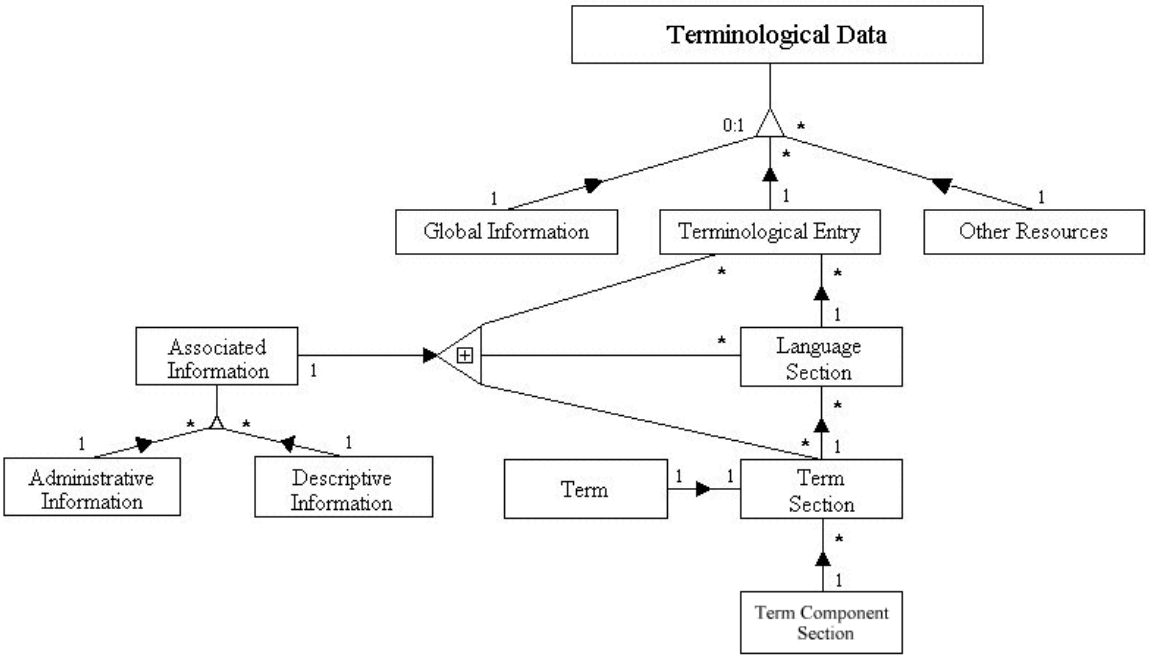
For covering multilingual aspects it is intended to use EuroWordNet for the translation of Ontologies (Declerck et al 2006). The ESPERONTO project/platform is based on EuroWordNet for ontology-based machine translation, using Wikipedia for multilingual content (parallel texts) and using linguistic enrichment of ontologies (Pazienza 2006).

As far as existing terminological, lexical, and semantic resources in the multi-risk domain cluster are concerned, the task in this prototype is to "ontologize" these resources.

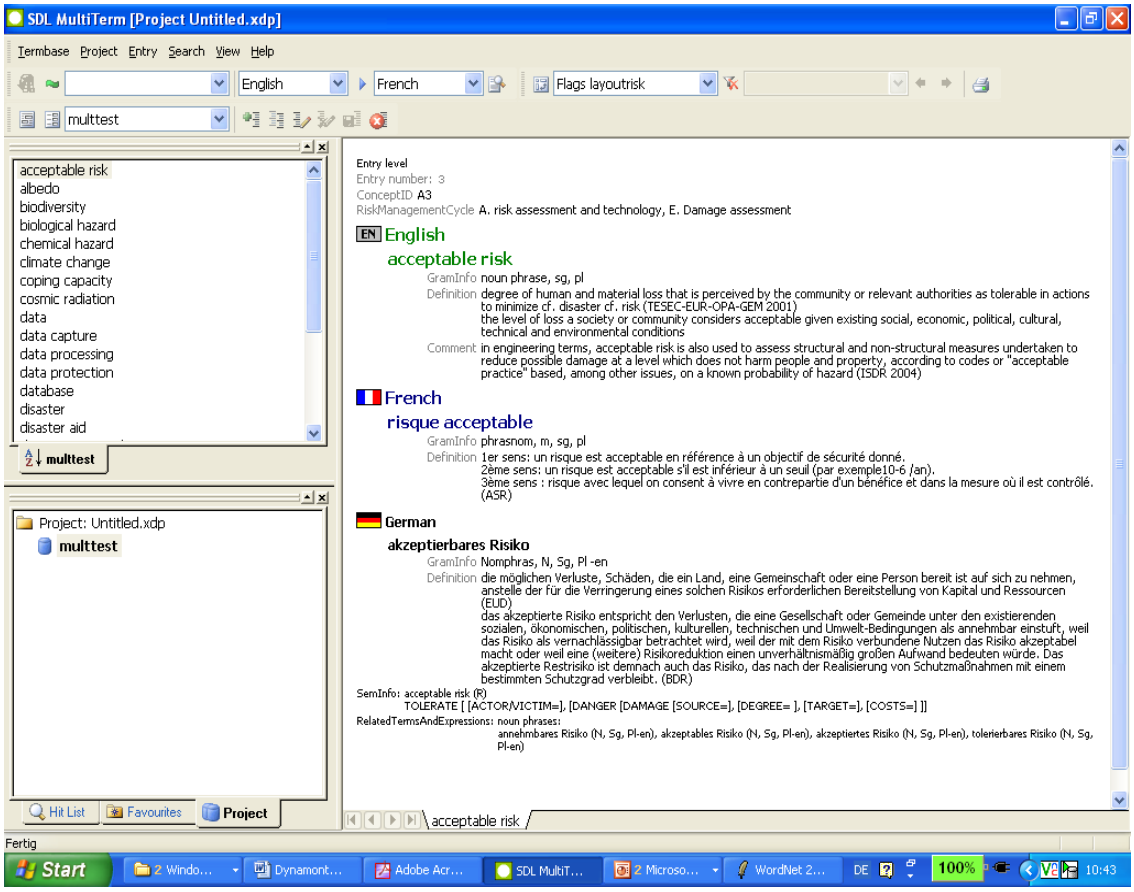
This includes the semantic enrichment of terminological resources as prepared in D2201, D2202, D2203, D2204. One point of departure is the existing multilingual risk glossary as it is currently modelled and represented in the online terminological database. The data model is based on ISO standards, i.e. the Terminology Markup Framework ISO 16642, the Data Category Registry Standard ISO 12620 and the industry standard Termbase Exchange Format (TBX) as an instantiation of ISO 16642.

The following figure shows the structural levels of terminological entries with their language sections, term sections, and term-component sections as the main body of xml documents, and with global information such as xml header and with other resources such as media files attached to terminological entries. We also distinguish different kinds of information resources such as administrative information and descriptive information. Each information resource type is associated

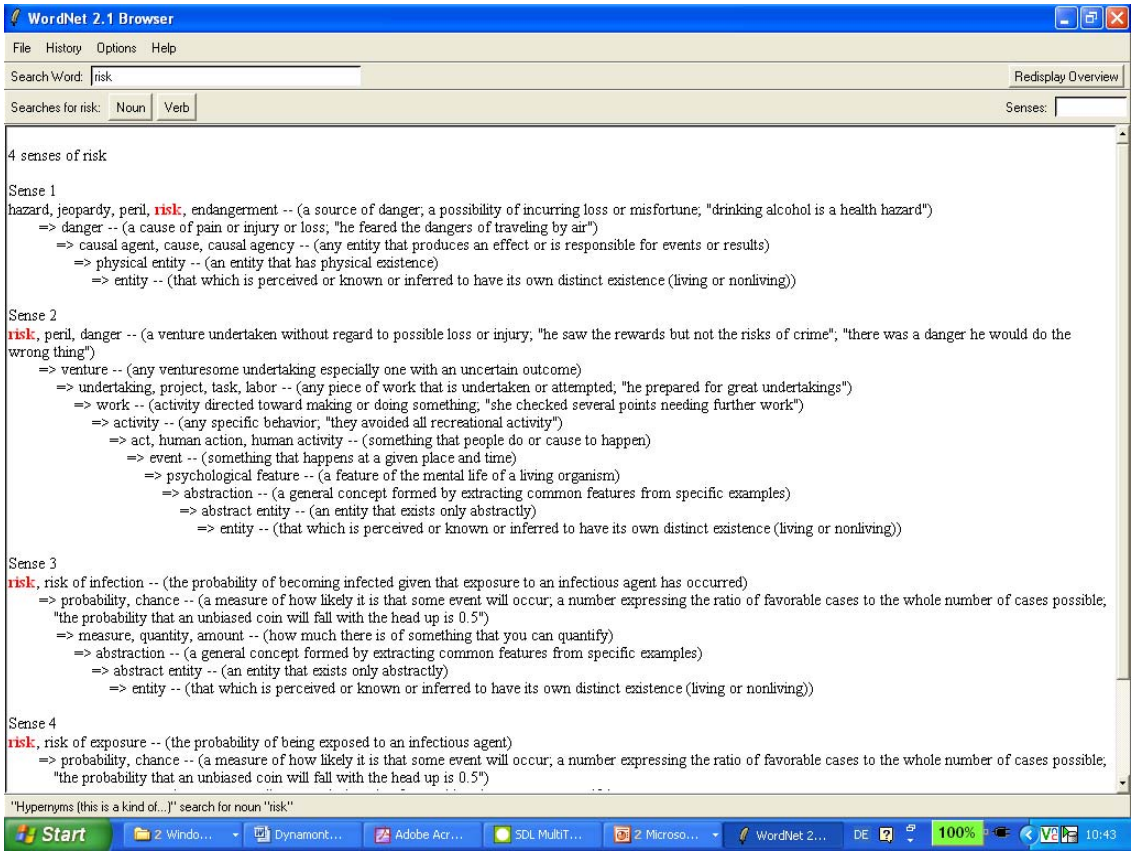
to many different kinds data categories that are standardized in ISO 12620 (that is organized and expressed in RDF).



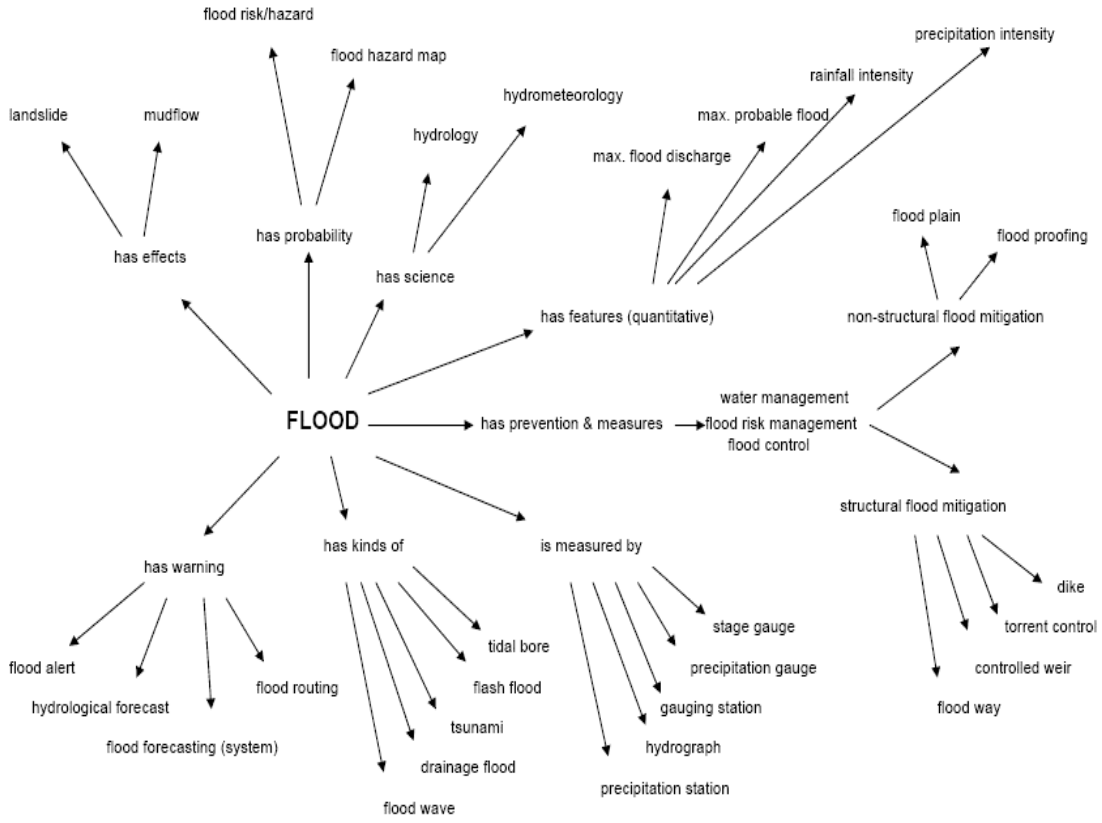
The following figure is screen-shot from the termbase that operationalizes the abstract data model:



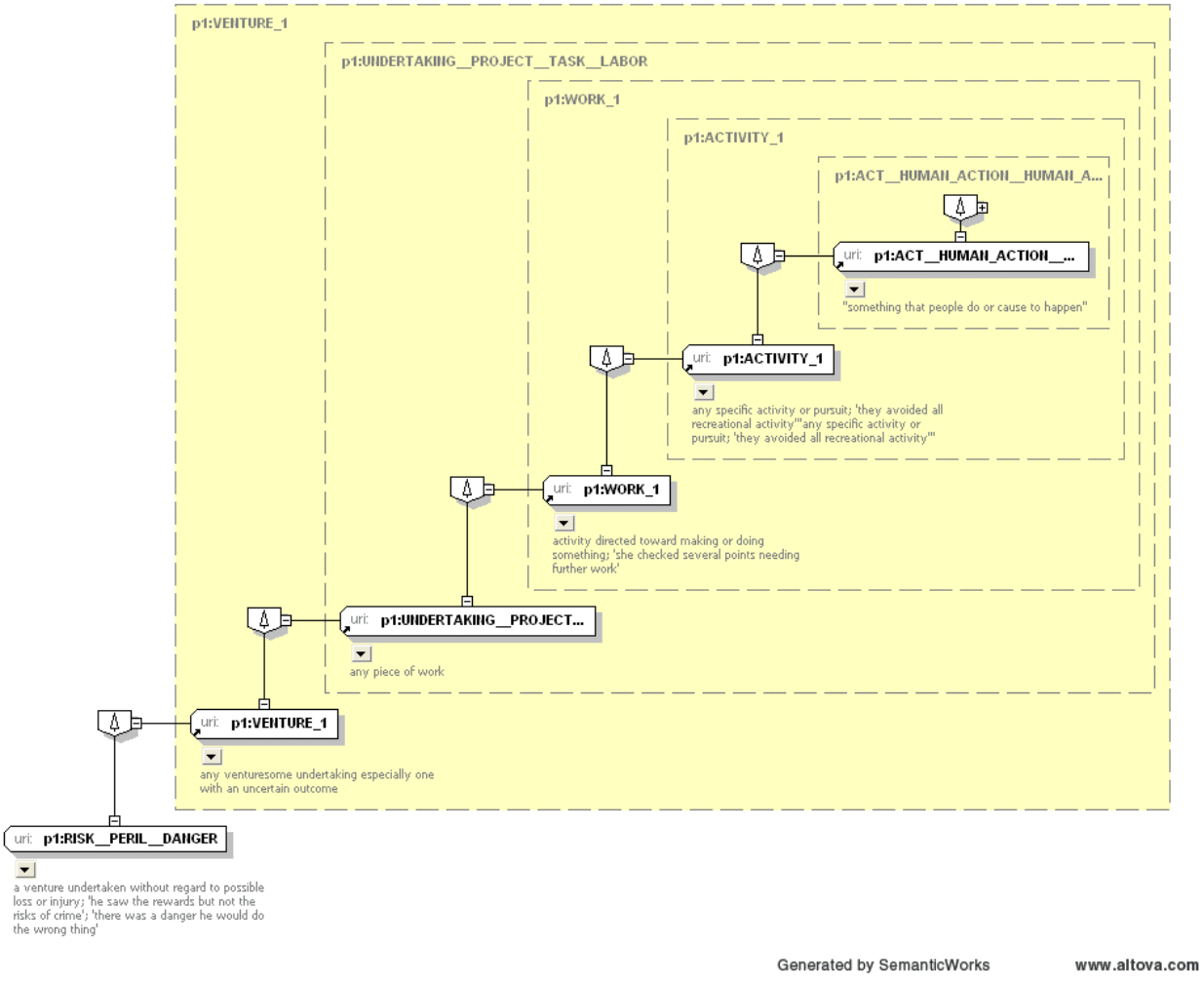
The following screen-shot shows an entry from the WordNet database on the word "risk". It shows the ontological foundation from the SUMO foundational ontology as it has been integrated into WordNet as a categorial grounding for words of the general lexicon.



The ontologization work also includes the explicit specification of conceptual links. This is an essential step in semantic enrichment and generates key components of terminological information for building full-fledged ontologies. The following example from the risk domain shows this step as part of not-yet fully formalized terminological work, obviously the next logical step in the workflow is to formalize the conceptual system. This can be done in different ways and with different methods (ranging from UML class diagrams, OWL class hierarchies, SKOS models, to frame-semantic representations (FrameNet). The operational step is to convert TBX documents (XML) to SKOS documents (RDF) and from there to OWL formalism. FrameNet documents are also mapped to OWL.



The next step in the workflow is the OWL representation in an ontology editor, in this case Altova Semantic Works, using the risk concept example:



With this fully formal ontology representation it is then possible to support WIN service architectures with different types of ontologies such as service and task ontologies.

In concluding we can summarize that all necessary methodological building blocks with their theoretical justifications have been created for use in other work packages of the WIN project.

7 BIBLIOGRAPHY

Budin, G. (2006). L'apport de la philosophie autrichienne au développement de la théorie de la terminologie: épistémologie, ontologie et théorie de l'objet. In: Langage (2006), no. 2 (in press)

Budin, G. / Melby, A. (2000). Accessibility of Multilingual Terminological Resources - Current Problems and Prospects for the Future. In: Proceedings of the LREC Conference – Language Resources and Evaluation, Athens, June 2000, vol. II, S. 837ff

Daconta, M./ Smith, K./ Obrst, L. (2003). The Semantic Web: The Future of XML, Web Services, and Knowledge Management. Wiley

Declerck, T et al (2006). Using EuroWordNet for the Translation of Ontologies. OntoLex Workshop Genoa 2006

Embley, D. Kurtz. B., and Woodfield, S., 1992. *Object-oriented Systems Analysis: a Model-driven Approach*. New Jersey: Prentice Hall

Gómez-Pérez, A./ Fernández-López, M./ Corcho. O. (2003). Ontological Engineering. Springer Verlag

Greciano, G. / Budin, G. (2006). Designing Linguistic Support for Risk Management Communication. In: EU-MEDIN Handbook (in preparation)

Guarino, N. / Giaretta, P. (1995). Ontologies and Knowledge Bases. Towards a Terminological Clarification. KBKS 1995

ISO 12 200: 1999 *Computer-Applications in Terminology – Machine-readable Terminology Interchange Format (MARTIF) – Negotiated Interchange*. Geneva: International Organization for Standardization

ISO 12 620: 1999 *Computer Applications in Terminology – Data Categories*. Geneva: International Organization for Standardization

ISO 16642: 2003 Terminology Markup Framework (TMF). Geneva: ISO

Madsen, B.N. et al. (2001) Defining Semantic Relations for OntoQuery. In: Per Anker Jensen & Peter Skadhauge (eds.): Proceedings of the First International OntoQuery Workshop, January 17-18. University of Southern Denmark, Department of Business Communication and Information Science, 2001, s. 57-88.

Maedche, A. (2002). Ontology Learning for the Semantic Web. Kluwer

Obrst, L. (2003). Ontologies: Definitions, Issues, & Problems. Santa Fe, Open Forum on Metadata Registries, 2003

Obrst, L./ Liu, H. (2003). Knowledge Representation, Ontological Engineering, and Topic Maps. In: Jack Park, Sam Hunting (eds). XML Topic Maps. Creating and Using Topic Maps for the Web. Addison-Wesley 2003, p. 103-148

Pazienza, M.T. (2006). Linguistic Enrichtment of Ontologies: a methodological framework. OntoLex workshop Genoa 2006

Peirce, C.S. (1991). Peirce on Signs. Writings on Semiotic by Charles Saunders Peirce. Edited by James Hoopes. (184-189). Chapel Hill and London: The University of North Carolina Press

Scheffczyk, J. / Baker, C.F. / Narayanan, S. (2006). Ontology-based Reasoning about Lexical Resources. Berkeley

Sowa, J. (2000). Knowledge Representation. Logical, Philosophical and Computational Foundations. Brooks/Cole, Pacific Grove

END OF DOCUMENT