UNIVERSITY OF LATVIA
FACULTY OF COMPUTING

Andrejs Vasiļjevs

Consolidation of Heterogeneous Terminology Resources

Doctoral Thesis

for Ph.D. (Dr. sc. comp.) academic degree

Scientific Advisor

Dr. habil. sc. comp., Professor
Juris Borzovs

RIGA, 2010

ABSTRACT

This thesis addresses the issues and solutions involved when consolidating heterogeneous multilingual terminology resources that are dispersed throughout numerous collections, publications and databases to provide single access point for both human users and web-services. Online availability of consolidated terminology resources from diverse sources is of utmost importance in translation practice and domain specific communications. One of the major goals for consolidation is to provide a single unified web-based access to distributed multilingual terminology resources. Unified methodology has been developed covering all major aspects from scenario based requirements analysis to data modeling, data storage, exchange and representation. The federation approach proposed in this work allows the consolidation of various existing terminology databases and centrally stored resources. This thesis introduces a new concept of terminology entry compounding for identification and unification of matching multilingual entries from different collections. Application of international standards is discussed to ensure global interoperability of terminology resources and integration into global language resource infrastructure. The practical results from using these approaches in the development of the EuroTermBank terminology databank are described. For the first time heterogeneous multilingual terminology resources are integrated a[IS1]nd database federation is established with a unified online interface, serving as a prove-of-concept for the approaches described in this work.

ACKNOWLEDGEMENTS

The author wishes to express sincere gratitude to his supervisor Professor Juris Borzovs who encouraged him to start PhD studies and wisely guided him through the completion of this research work.

The preparation of this thesis work was supported by European Social Fund. The thesis is based on the research and development project for EuroTermBank terminology databank. As the coordinator of the EuroTermBank project the author would like to acknowledge the support for the project from the European Union eContent Programme. Research on terminology integration into desktop authoring software was supported by the European Union EUREKA Programme.

The author would like to express many thanks to colleagues from the EuroTermBank project partner organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland).

Personal thanks to Professor Klaus Dirk-Schmitz from Cologne University of Applied Science for his skillful support of the management of international team of researchers.

The author is grateful to Professor Valentīna Skujiņa and Dr. Ilze Ilziņa for introducing him to the field of terminology and raising awareness for research and IT solutions development for terminology management.

The author would like to thank his colleagues and co-authors of several scientific papers Raivis Skadiņš, Signe Rirdance, Andris Liedskalniņš and acknowledge their valuable contribution in the research and development work related to terminology consolidation in EuroTermBank.

Table of Contents

1. INTRODUCTION

Information technologies are transforming almost every field of human activity and terminology is of no exception. Eugen Wüster, regarded as a founder of modern terminology, claimed that computer science is one of the keys to terminology because of the enormous possibilities it offers to store and retrieve information and to order conceptual systems (Wüster, 1968).

This work provides a research in the application of the state of the art in computer science and information technology for addressing one of the major problems confronting the terminology field − consolidation of heterogonous terminology resources.

The following chapter briefly introduces the research area, describes the motivation for and aims of the research. The research hypothesis stated by the author is proved both by theoretical research work and practical implementation through the EuroTermbank project. The key results of the research are listed and the author’s contribution is specified. At the end of this section an outline of the remainder of the thesis work is given.

1.1 RESEARCH AREA

Consistent, harmonized and easily accessible terminology is an extremely important stronghold for ensuring true multilingualism in the European Union and throughout the world. From legislation and trade to the needs and mobility of every EU citizen, terminology is the key for easy, fast and reliable communications. Uniform terminology enables to ensure that the same meaning is conveyed between participants in a written or oral communication. This puts terminology work in a crucial role for unambiguous and reliable communication. The rapid path of changes in many technological and economical areas leads to an ever growing introduction of new concepts and terms to describe them. Efficient and reliable communication in specialized areas depends on the efficiency of introducing, disseminating, and applying these new terms in practical use.

Historically, most terminology resources have been developed within a rather narrow setting of an organization, a company or an industry sector, very often related to translation needs. This has resulted in the fragmentation of resources across terminology holders and the limited availability of harmonized terminology data on the national and supranational levels (Henriksen et al., 2006). Despite the fact that international standards have been developed, a wide proliferation of data models and technical formats, including proprietary ones, is a given, and adoption of existing standards and recommendations has been rather slow.

Globalization from the one side and growing language awareness from the other side dictate the need to consolidate terminology resources, harmonize international terminology, and provide online access to reliable multilingual terminology. Demand for the creation of consolidated multilingual terminology resources is growing, both in the public sector, as governments are required to communicate with their citizens in more and more languages, and in the commercial sector, as companies move to communicate with their customers in multiple languages simultaneously across the globe.

Advances in language technologies and machine translation are about to change the traditional patterns of the creation and use of language resources. New approaches and platforms are urgently required to support these requirements.

Response to this demand for new models of terminology consolidation and distribution requires the application of state of the art in information system development.

1.2 MOTIVATION OF THE RESEARCH

The overall situation in terminology is characterized by many gaps and problems. The major developers of terminology include public institutions, universities and technical societies as well as representatives of the private sector. Although there are a significant number of institutions involved in terminology work, very few of them produce resources that are exchangeable or marketable.

Insufficient distribution and reutilization of existing resources has long been identified as one of the major weaknesses in the European terminology landscape (Ahmad et al., 1996). This situation was attributed to, among other obstacles, the lack of information and awareness, generally low level of technological support for terminologists and lack of standard data interchange formats.

In many EU countries there is a lack of coordination between institutions dealing with terminological activities. This often results in useless efforts or duplicate results (COTSOES, 2002). Terminologists and subject specialists have little contact with their colleagues working on similar subject areas. Across subject fields and in different sectors potential users of terminology are often not even aware of the resources available~~-~~.

The reasons for this lack of communication are a general fragmentation of the creation and distribution mechanisms on the institutional, sector/industry and national levels (Ahmad et al., 1996). Term banks tend to be small in size, mostly highly specialized, difficult to access. These difficulties are amplified by considerations of confidentiality, institutional restrictions and legal uncertainty about copyright status of particular terminology resources.

As a result, there is a lack of easily accessible terminology resources and the existing resources are not adequately reutilized. The quality of available terminological collections varies widely and is inadequate in many cases. International standards are not always used in terminology development and sometimes even unknown to the people directly involved.

Fragmentation of terminology resources is particularly acute in the new European Union member countries that have undergone rapid social and economic transformations and urgently need to integrate their terminology development with the rest of the EU and the global economy. At the beginning of the research work new EU member countries faced the issue of terminology resource fragmentation across different institutions, inconsistency and lack of coordination in terminology development, as well as structural and technical incompatibility (EuroTermBank, 2005). Rapid development and dissemination of new terms is especially important for smaller languages. Placing terminology work and implementing terminology consolidation in widely accessible databases is among key tasks of national language policies in several countries (Auksoriūtė, Gaivenytė, & Umbrasas, 2003; IZM, 2006; Vasiļjevs, 2008; Thelen & Steurs, 2010).

A great deal of terminology data is available only in the form of printed dictionaries and bulletins or stored in card files. The transformation from centralized terminology development during the Soviet era with the focus on the Russian language to the requirements of market economies was not fully completed. This has led to the lack of coordination between the institutions involved in terminology development, inconsistency and poor quality of terminology data, and insufficient mechanisms for the dissemination of new terminology.

1.3 THE AIM OF THE RESEARCH

In our research we focus on the problem of consolidation of heterogeneous multilingual terminology resources. The heterogeneous nature of these resources is characterized by different data structure, language coverage, organization principles, formatting, storage formats, and geographical location.

This work describes the main issues related to terminology data harmonization, collection, and access. The main challenge is to aggregate resources convergingfrom many terminological data sources with varying structures and to present them to the user as one consistent source of terminological information.

For his research author has established the following hypothesis:

Access and usability problems posed by the fragmentation and heterogeneity of terminology resources can be effectively solved by a federated multilingual terminology portal that provides consolidated data representation and is integrated in authoring software.

The goal of this research is to create a unified methodology that encompasses all the major aspects related to the consolidation problem:

- Requirements analysis in a multinational multi-actor and multiuser environment;
- Data modeling principles for terminological information;
- Data storage and data exchange mechanisms;
- Consolidation approach for independently maintained terminology databases;
- Unified representation of dispersed heterogeneous terminology data.

1.4 PRACTICAL IMPLEMENTATION OF RESEARCH RESULTS

The research activities of this thesis work are closely related to the EuroTermBank project. Part of the research was carried out in the framework of EuroTermBank together with an international team of scientists and IT practitioners from EuroTermBank Consortium.

Successful implementation of the proposed methodology in EuroTermBank system serves as the proof of research hypothesis and methodology described in this work.

EuroTermBank project is targeted at facilitating terminology data accessibility and exchange with a goal to collect, consolidate and disseminate dispersed terminology resources through an online terminology data bank (Rirdance & Vasiljevs, 2006). EuroTermBank is part of the European Union eContent Programme, which is aimed at promoting European internet resources and multilingualism.

EuroTermBank project was initiated and led by the author of this thesis work. The project was carried out by 8 partners from 7 European Union countries – Germany, Denmark, Latvia, Lithuania, Estonia, Poland and Hungary. The project partners are Tilde (Latvia), Institute for Information Management at the University of Applied Science Cologne (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland).

The project was part of the European Union eContent program aimed to facilitate the production, use and distribution of European digital content and to promote linguistic and cultural diversity on the global networks.

The initial focus of the EuroTermBank was to contribute to the improvement of the terminology infrastructure in selected new European Union member countries (Latvia, Lithuania, Estonia, Poland, Hungary), however EuroTermBank continues to expand its activities to other countries in the EU and beyond. This aim is accomplished by establishing cooperative networks of terminology institutions on various levels and by consolidating and harmonizing existing terminology resources.

EuroTermBank enables the exchange of terminology data with existing national and EU terminology databases by establishing cooperative relationships, aligning methodologies and standards, and designing and implementing data exchange mechanisms and procedures. Through harmonization, collection and dissemination of public terminology resources, EuroTermBank is aimed to facilitate enhancement of public sector information and strengthen the linguistic infrastructure in the new EU member countries.

Development, population and maintenance of a web-based terminology data bank constitute the major tangible outcome of the project. The data bank works on a two-tier principle – as a central database and as an interlink node or a gateway to other national and international terminology banks.

The project outcome is a reliable multilingual terminology resource, networked with other existing national and international resources available for users over the global internet community.

1.5 KEY RESULTS OF THE RESEARCH

To solve the key research objective – consolidation of heterogeneous terminology resources, the author has reached the following key results:

- Methodology for consolidation has been developed including all crucial steps in consolidation of resources;
- Requirements analysis method is proposed based on terminology work scenarios. Three distinctive scenarios are identified – local, national and international scenarios;
- Data modeling guidelines for terminology work scenarios are provided;
- Data storage and exchange standards are analyzed, applicability of TBX for data storage is proposed and demonstrated experimentally;
- Federation principle is proposed and implemented for consolidation of independently maintained terminology databases;
- Terminology entry compounding mechanism is introduced for consolidated representation of terminology data;
- Corpus based analysis methods are suggested and experimentally affirmed for terminology entry compounding;
- Proposed methods are proved by practical implementation in EuroTermBank project - largest online source of terminology in multiple subject fields in

languages of new European Union countries is developed containing about 2 mil. term entries.

Research results were achieved by an international team under the author's leadership. The author has made the specific following contributions:

- Problem of terminology resource fragmentation defined;
- Unified methodology for terminology consolidation proposed and elaborated;
- Federated database approach proposed and implementation ensured;
- New method for consolidated term representation – term entry compounding – proposed and elaborated;
- EuroTermBank project initiated, international researcher and development team formed, project leadership in implementation of proposed methods provided.

1.6 AUTHOR'S PUBLICATIONS AND PRESENTATIONS RELATED TO THE RESEARCH

The author has presented the results of the research at 21 international conferences, workshops and seminars:

- Terminology and resource harmonization, PhD level course by the Marie Curie CLARA project, September 2010;
- Terminology and Knowledge Engineering Conference TKE 2010, Dublin, August 2010;
- The 7th International Conference on Language Resources and Evaluation LREC 2010, Malta, May 2010;
- International Terminology seminar of the Network to Promote Linguistic Diversity (NPLD), Dublin, December 2009;
- The Fifth Conference of the EUREKA National Coordinators from Nordic and Baltic Countries, July 2009;
- tcWorld Conference 2008, Wiesbaden, November 2008;
- The Eight International Conference on Terminology and Knowledge Engineering TKE 2008, Copenhagen, August 2008;
- LREC-2008 Workshop on Uses and usage of language resource-related standards, Marrakech, May 2008;

- The Third Language and Speech Technology Conference LangTech 2008, Rome, February 2008;
- The First International Conference on Global Interoperability for Language Resources ICGL 2008, Hong Kong, January 2008;
- The Third Baltic Conference on Human Language Technologies, Kaunas, October 2007;
- International Conference Recent Advances in Natural Language Processing, Borovets, Bulgaria, September 2007;
- Seminar by EuroTermBank Consortium ”Towards Consolidation of European Terminology Resources”, Luxembourg, March 2007;
- International conference on terminology issues “The impact of terminology on everyday life”, Antwerp, November 2006;
- EAFT Third Terminology Summit, Brussels, November 2006;
- International Conference “Terminology of national languages and globalization”, Vilnius, October 2006;
- The Third International Conference on Terminology, Standardization and Technology Transfer, Beijing, August 2006;
- The Seventh International Baltic Conference on Databases and Information Systems, Vilnius, July 2006;
- LREC 2006, the 5th International Conference on Language Resources and Evaluation, Genoa, May 2006;
- The Second Baltic Conference on Human Language Technologies, Tallinn, 2005;
- The First Baltic Conference ”Human Language Technologies – the Baltic Perspective”, Riga, April 2004.

Research results are reported in the 14 [IS2]papers published in the proceedings of the international conferences, one journal article and one collective monograph co-edited by author (see list of author’s publications on page 110).

2. BACKGROUND AND RELATED WORK

In last three decades terminology and computer science have become intrinsically connected. These disciplines have mutually beneficial relationships – computer science assists and changes terminological activities and its methodology, and terminology helps research in computational linguistics (Cabré, 1999).

2.1 INTRODUCTION TO TERMINOLOGY FIELD

The term *terminology* somehow confusingly is being used in two senses – to denote the scientific discipline of terminology and to denote the set of terms from a discrete subject field (Wright & Budin, 2001).

Terminology as a scientific discipline studies the structure, formation, development, usage and management of terminologies in various subject fields (ISO 22128, 2008). To avoid possible confusion, terminology as a discipline in some publications is also called *terminology work* (ISO 704, 2009).

The set of designations of concepts belonging to one special language is also called terminology (ISO 1087-1, 2000). S*pecial language* is defined as a language used in a field of specialized knowledge and characterized by the use of a specific linguistic means of expression.

The basic principles and methods of terminology work are defined by (ISO 704, 2009)[IS3]. They are rooted in the so called Vienna School of Terminology established by Eugen Wüster (Felber, 1984) and his *General Theory of Terminology*[IS4] (Wüster, 1972). It should be noted that in recent years interest has revived to revise this theory and some alternative theories have emerged, like socioterminology and Communicative Theory of Terminology (Cabré, 2003). Still the Vienna School dominates the field and is the most established, elaborated and widely accepted. For this reason the author will abide by it in this thesis work.

According to these principles any terminology work starts with concepts. In terminology work, *concept* is a unit of knowledge corresponding to the class of objects. *Object* is defined as anything perceived or conceived. Some objects, such as a car, a tree, or a table, can be considered concrete or material; others, such as gross domestic product, gravity, or inflation, can be considered immaterial or abstract. Not every individual object is differentiated and uniquely named. Instead through the process of abstraction, individual objects are conceptualized into units of knowledge called *concepts*. For terminology work, concepts are considered mental representations of objects within a specialized context or field.

Terminology work identifies concepts of a particular subject field and assigns designations to these concepts. Concept designation is usually one or more words and is called a *term*. Concept designations can also be non-lexical such as formulas, codes, symbols, visual depictions or audio signals.

(Cabré, 1999) defines *term* as a lexical unit with a morphological or a syntactic structure which corresponds to a minimal autonomous conceptual unit in a given field.

One of the major goals of terminology work is to ensure precise and accurate communication in a given field. For this purpose ideally in a given subject field a given term should be attributed to only one concept and that given concept is represented by only one term. Such a one-to-one relationship between concept and term is called *monosemy*. One-to-many and many-to-one relationships between terms and concepts are called *homonymy* and *synonymy*, respectively. Such occurrences can lead to ambiguity. To avoid this in *prescriptive* terminology only one term – called *preferred term* – should be used by subject specialists (ISO 704, 2009).

Our thesis is related to terminology management that is understood as any deliberate manipulation of terminological information (Wright & Budin, 2001).

We are dealing with *terminology resources* - sets of terms from a particular subject field that are documented or recorded in some information medium.

(Ahmad et al., 1996) states that "terminological resources are valuable in many ways: as collections of names or other representations, as the object of standardization and harmonization activities, and as the input (or output) of a wide range of applications and disciplines, whether human or machine-based".

Concepts are represented not only by terms but also by definitions. A terminological definition is a concise description of the delimiting characteristics of a concept, presented in lexicographical, or dictionary-like, format. The definition must give the meaning of the term, rather than dealing with questions of the term's usage (Pavel & Nolet, 2001).

In terminology work, to model a concept system, the concepts of the concept field have to be examined and compared. As a minimum the following relations shall be used to model a concept system (ISO 704, 2009):

- hierarchical relations;
- generic relations;
- partitive relations;
- associative relations.

Uniform terminology enables to ensure that the same meaning is conveyed between participants in a written or oral communication. This puts terminology work in a crucial role for unambiguous and reliable communication.

2.2 TERMINOLOGY DATABASES

The enormous potential that application of computer technologies and computer science can bring for terminology development was recognized already by Eugen Wüster who is regarded by many as a founder of modern terminology (Wüster, 1968). Since then different software systems have been developed to process term related information collectively regarded as terminology management systems.

A terminology management system is a software tool specifically designed for collecting, maintaining, and accessing terminological data (ISO 26162, 2010). It can be designed and built to process terminological data in a dedicated way or integrated into other kinds of application software (UNESCO, 2005).

Terminological data is organized into *terminological entries* (short form: term entries) or terminological records. A terminological entry treats a single concept and contains all terminological data related to that concept. Among other terminological data it contains all the terms designating that concept either in one or multiple languages (ISO 26162, 2010; Wright & Budin, 2001).

A terminological resource (also called terminological data collection) is a text or data resource consisting of terminological entries. Usually it contains terminological data about concepts from a particular subject field in one or multiple languages. A terminological entry is part of a terminological resource that contains the terminological data related to one concept.

(ISO 26162, 2010) defines a terminological database or termbase as a database comprising a terminological resource. In literature terminology systems dealing with structured data collections are called either *terminology databases* (short form: *termbases*) or *terminological data banks* (short form: *term banks*). Distinction between these two categories is not strict and is drawn mostly according to size, complexity and application scope of the terminology system.

(Wright & Budin, 2001) distinguishes termbases from term banks as being individual databases that are frequently produced by individuals, companies, agencies, etc.

Term banks in contrast usually a constitute wider spectrum of institutional resources which are made accessible to wider groups of users, in many cases on a subscription basis. (Sager & McNaught, 1980) define a terminological data bank as a collection of special language vocabularies, including standardized terms, stored in a computer which can be used as a mono-lingual or a multilingual dictionary for direct consultations, as a basis for dictionary production, as a control instrument for consistency of terminology and as an ancillary tool in information and documentation. Term banks frequently require public funding.

In our view these attempts to distinguish termbases and term banks into two distinct classes are not fully motivated. Instead of being two separate classes of systems, in our view term banks are rather a subclass of a broader class of terminology databases.

This broader view is supported by (UNESCO, 2005) providing a more general description of termbases as systems containing mono- or multilingual terminological data which can be established at country, language community or local level depending on the needs of the respective communities.

According to (Sager, 1990) a terminological data bank can be viewed as a set of special language vocabularies with the following characteristics:

- The information in stored in computer systems;
- They include nomenclatures, special terms and phrases, with the information necessary for their identification;
- They can be used as monolingual, bilingual or multilingual dictionaries;

- They offer on-line access;
- They are the basis for dictionary production;
- They are used to monitor the vitality of a language and the creation of terms;
- They are ancillary tools for information and documentation.

The distinctive characteristics of term banks are multilinguality, multi-disciplinarity, provision of multiple terminological resources, and different user groups from different institutions. Nowadays many term banks provide an online interface and are accessible for free or on a subscription basis to any interested user.

(Felber, 1984) recommends the following steps for a terminology project that can be applied to both manual and computerized handling of data:

Step 1: Define the field of study (the subject field).
Step 2: Decide on the structure of terminological data.
The most important data to be included is:

- Date of record indicating the recorder;
- Serial number;
- Classification symbol for the place in system of concepts;
- Terms(s) designating the concept;
- Synonymous terms;
- An explanation of the concept (definition);
- Term in context (example of usage);
- Illustration;
- Authority and country symbols;
- Language symbol;
- Explanatory notes
- Term designating the broader concept;
- Term(s) designating concept(s) of the same abstraction level;
- Term(s) designating the narrower concept(s);
- Sources, where terms, definitions, illustrations have been found;
- Code symbol for the volume (for card based term record storage).

Step 3: Choosing languages for multilingual data collections.

Step 4: Deciding on symbols to be used for lexicographical information related to terms.

Step 5: Guidelines should be prepared for principles and methods in terminology work preparing terminological data. These guidelines should correspond to ISO standards.

In (Cabré, 1999) the following steps are specified in planning the process for the creation of a term bank:

Step 1: Definition of the major features expected of the term bank. This should be done through the identification of user needs:

- Identification of target users;
- Delimitation of the needs of each user group;
- Comparison, coordination, prioritization of the needs identified.

Step 2: Feasibility study including aims and functions of the bank, hardware and software requirements, data size and organization, major channels of dissemination and financial model.

Step 3: Basic design of the data bank covering the description of the overall structure of the term bank and the processing of the data. Include general features, file and data conversion procedures, system compatibility, size criteria, presentation of information to users and other system design aspects.

Step 4: Chose between proprietary software development and adaptation of commercially available software.

Step 5: Detailed design of the term bank including decisions regarding:

- Entry: type of information, sources of the data, entry system, organization of the entry, structure of the information, entry protocol, etc.
- Storage: type of records, relationships among the records, structure of the records, protocols of representation and use of records, etc.
- Retrieval: types of queries that are to be answered, ways of retrieval, formats for retrieved information, typology of users, etc.

Step 6: Implementation of the term bank. Prototype development or pilot project may precede implementation of the full system.

Step 7: Periodic reviews and updates of the term bank to meet evolving user needs.

(Cabré, 1999) stresses the importance of involving representatives of real users to ensure that the term bank will meet the actual needs of users.

2.3 FIRST TERMINOLOGY DATA BANKS

Several institutions dealing with terminology recognized the potential of computerization of terminological data already in 1960-ies. This recognition materialized in the first terminology databases that became one the first large scale databases of linguistic data.

The earliest term banks were developed in the mid-1960s and early 1970s by translation departments in large organizations. Their main functions were to supplement printed dictionaries by providing up-to-date multilingual terminology, to make terminology produced in-house more widely available, to facilitate the implementation of consistent and unified terminology among different translators, to speed up the translation process, and to serve as instruments for language planning and standardization (Kent, 1998).

The first large scale attempt to computerize terminological vocabularies goes back to the 1964 European Commission projects DICAUTOM and EUROTERM. Later, in 1971 these attempts resulted in the Commission's terminology database EURODICAUTOM (Felber, 1984).

Projects on national databases for standardized terminology started in several countries. The main focus of these systems was on the organization of terminological data used in the prescriptive environment for standardization and normative purposes. The NORMATERM database in France was created in 1973 (Laurent, 1977). It was developed by the French standardization institution AFNOR (*Association Française de Normalisation*). AFNOR decided to list, classify and make available the terminological information, which hitherto had been scattered amongst all the French standards. NORMATERM system was designed to assist users in standardization of related tasks - to provide assistance in compiling the ISO Standardization thesaurus, to enable retrieval of standardized terms, to provide indexes of terms, to process standardized terminology in the French and international standards.

A somewhat similar database, DIN-TERM, was developed in 1976 to serve the terminology needs in the standardization process in the Federal Republic of Germany (Felber, 1984). In its current version the database contains the standard terminology laid down in the German standards (DIN, DIN EN and DIN EN ISO Standards and draft Standards) and from the International Electrotechnical Vocabulary.

In 1974 the Soviet terminology bank by GOSSTANDART was launched: *Справочный банк терминов – автоматизированная система информационно-терминологического обслуживания (СБТ АСИСТО)*[1].

The TEAM system by Siemens AG was first started in 1967. It consists of lexical entries based on a defined concept and offering terms expressing this concept in up to eight languages – German, English, French, Spanish, Russian, Italian, Portuguese, and Dutch (Brinkmann, 1980; Schulz, 1980). TEAM was among the first to realize the need for and benefit to be gained from serving different types of users, involving many partners active in contributing terminology in many fields, and providing diversified services, catering for translators, publishers, standardization specialists, information scientists and language teachers (McNaught, 1993).

In Canada the TERMIUM database was first created at the University of Montreal and later become the database of the General Division of Terminology and Documentation (Dubuc, 1972). Other examples of early success in large scale terminology systems are LEXIS, first released in 1966, by *Bundessprachenamt* in Germany and TERMDOK, launched by TNS, in Sweden in 1968.

The first term banks, including EURODICAUTOM, Termium, TEAM, LEXIS, were mostly term-oriented. The terminological data was structured around a term as a lexical unit assigning all possible meanings to a particular term.

The second generation of term banks started to implement a concept-oriented approach, where concept is in the center of terminological data organization. Here the lexical unit term is subordinated to a concept-based entry defined by a definition, illustration or nomenclature code. Facilities for representing hierarchical relationships between concepts were provided. The Danish multidisciplinary term bank

[1] *Terminological data bank – Automated system of the terminological information service* (transl. from Russian).

DANTERM, the Norwegian term bank on oil terminology NoTe, and the medical term bank on virology SURVIT are examples of these second generation term banks.

According to the categorization suggested by (Nkwenti-Azeh, 1993) the so called third generation of term banks are knowledge-oriented. Terminology is viewed as a problem-oriented, specialized knowledge representation, and the terminological database can be seen as an expert system for terminology.

Terminology and editing tools should be integrated in the so-called translator workstations or translators workbenches. Those should provide access to external term banks and provide automatic identification of terms (Nkwenti-Azeh, 1993).

2.4 TYPOLOGY OF TERMINOLOGY BANKS

Most of the data banks that currently exist were designed to aid translation and usually contain terminological information from the lexicographic and terminological literature such as lexicons, dictionaries, encyclopedias, vocabularies, glossaries, etc. Their primary purpose is to facilitate translation by giving translators a one-stop, user-friendly tool for queries that includes several dictionaries and is capable of providing reliable suggestions for translation (Cabré, 1999).

Let us mention some of the types of term banks categorized by (Cabré, 1999) based on criteria suggested by (Sager, 1990) and (Felber, 1984):

A. Banks defined by objectives:
   - Informative banks, designed to disseminate terminology;
   - Prescriptive banks, designed to intervene in term usage.

B. Banks defined by their entries:
   - Banks based on terms;
   - Banks based on concepts.

C. Banks defined by subject matter:
   - Banks containing information about several subject fields;
   - Banks on a single special subject.

D. Banks defined by their size:
   - Large banks, usually of administrative bodies;

- Terminology minibanks, developed by a professional or a centre specializing in a subject field.

E. Banks defined according to the main interest of the data they contain:
- Term banks;
- Phrase banks;
- Banks of terms in documents (specialized texts);
- Encyclopedic banks (terms with encyclopedic information);
- Visual banks (images with captions).

F. Banks defined according to the choice of contents in relation to their objectives:
- Standard banks, containing only correct information;
- Descriptive banks, containing all types of information;
- Informative banks, containing all types of information but indicating their relationships to a standard.

G. Banks defined according to how the data is organized:
- Banks organized by document;
- Banks organized by terms without context.

H. Banks defined according to the hardware used – our adaptation to current situation (originally this category distinguish systems on mainframes and minicomputers):
- Server based termbases;
- Client based local termbases.

These criteria are not mutually exclusive and databases can have a mixture of these characteristics.

2.5 TERMINOLOGY CONSOLIDATION ON THE INTERNATIONAL LEVEL

Integration of multilingual terminology resources across countries has been addressed by several large projects in different international institutions. Every such effort has its own goals and conditions determining the resulting approach and solution. In the following chapter we will describe the consolidation of terminology used by European Union institutions in the IATE database, consolidation of terminology from international standards in the ISO/CDB database and the international consolidation of specific legal terminology in the LexALP system.

2.5.1 CONSOLIDATION OF EU INSTITUTIONAL TERMINOLOGY

A good example of terminology consolidation and harmonization on an international scale is IATE (Inter-Active Terminology for Europe), the EU inter-institutional terminology database (Johnson & Macphail, 2000; Rummel & Ball, 2001).

The EU institutions have been discussing the possibility of merging their terminology databases for many years. Practical steps started with a feasibility study.

There were three major EU terminology databases:

- Eurodicautom of the European Commission: about 5 million terms, Lenoch classification, web-based interface, integration with MultiTerm to support translation workflow;
- TIS of the European Council: about 600 000 terms, legacy classification with 170 subject codes, desktop application and web-based interface;
- Euterpe of the European Parliament: about 110 000 terms, legacy classification, integration with MultiTerm[IS5] to support translation workflow, web interface on intra- and internet.

Some smaller institutions (European Investment Bank, Court of Auditors, Translation Centre for the Bodies of the EU) had internal databases, generally using local TMS MultiTerm[IS6]. Others worked with glossaries in word processor formats, card files, etc. Certain institutions (European Social Committee/Committee of the Regions) have no systematic terminology arrangements.

The identified drawbacks were the lack of a single point of access to up-to-date terminology data for all the EU institutions, limited interactivity and as a result little user feedback, slow terminology cycle, inconsistency in the use of terminology between the institutions, no easy way of standardizing usage, difficult cooperation between terminology services of different institutions, resulting in a considerable duplication of effort.

The main recommendations of this study were:

- An inter-institutional database is both technically feasible and functionally desirable;
- All existing data should be merged into a single database;
- A common data model should be adopted;
- Common rules for data presentation and evaluation should be defined;
- Cooperative management mechanisms should be established;
- Full interactivity for data input and updating.

As the study clearly demonstrated the need for an inter-institutional consolidation of terminology, the IATE project was launched in 1999 with the following objectives:

- To provide a single point of access to all existing EU terminology resources;
- To provide an infrastructure for the constitution, shared management and dissemination of terminology resources;
- To provide a vehicle for the application of advanced language processing technology to terminology management;
- To provide a basis for integrating terminology into the translation and document workflow;
- To create a European platform for cooperation between EU institutions and terminology organizations in Member States.

IATE incorporated all of the existing terminology databases of the EU's translation services into a single interactive online inter-institutional database: EURODICAUTOM, TIS, Euterpe Euroterms (Translation Centre for the Bodies of the EU) and CDCTERM (European Court of Auditors).

To consolidate data it was necessary to deal with differences in database structure: the different philosophies of terminology and different historical backgrounds that were expressed in the data stored had to be reconciled. This process involved the definition of mapping rules between the data structures of the existing databases and the new format of the interinstitutional database. The data structure adopted a concept-oriented approach. The mono- and multilingual information on each aspect of a concept is expressed on four interrelated levels of the data structure of the terminological entries as illustrated in Figure 1 from (Rummel & Ball, 2001).

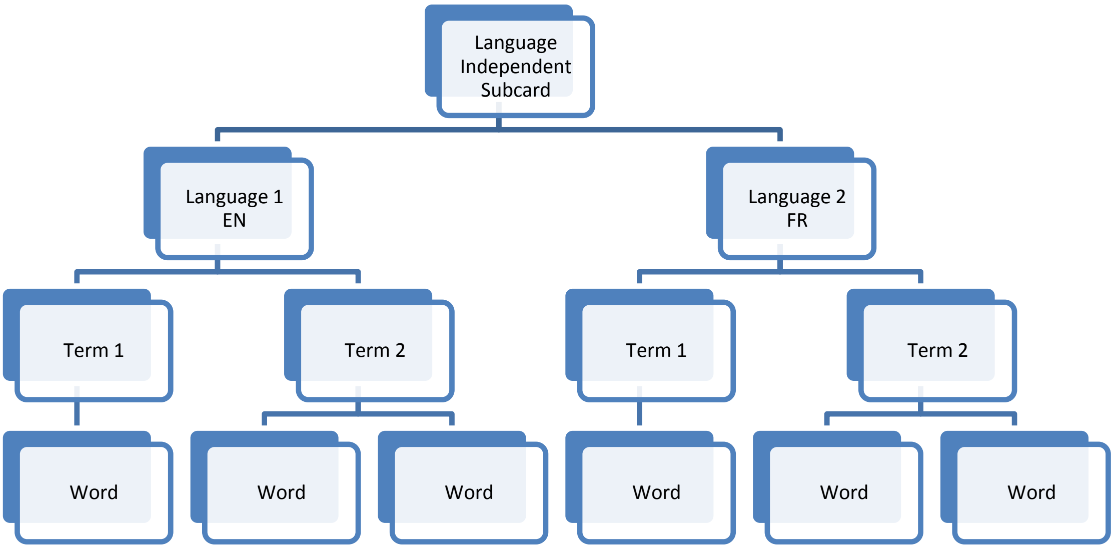


Figure 1 Basic data structure of the IATE termbase

For data mapping from legacy databases to IATE, manual rules and corrections were applied.

The following institutions of the European Union currently participate in IATE:

- European Commission
- European Council
- European Parliament
- European Court of Auditors
- Economic and Social Committee
- Committee of the Regions
- European Court of Justice
- Translation Centre for the Bodies of the EU
- European Investment Bank
- European Central Bank

Currently the IATE term bank contains about 1.4 million multilingual concept-based entries with 8.4 million terms, including approximately 540 000 abbreviations and 130 000 phrases. IATE covers terms in all 23 official EU languages.

Since the summer of 2004 IATE has been used internally by EU institutions and agencies for the collection, dissemination, and shared management of EU-specific terminology, and was released for online public access in 2007.

Analyzing current implementation of IATE we can conclude that it succeeds in providing a centralized system for all EU terminology resources as a single access point that serves the EU institutions as well as EU citizens (Rirdance & Vasiljevs, 2006).

At the same time we can see some limitations in fulfilling its initial broader ambition. IATE is a centralized database that is managed only by central EU institutions. There is no direct integration with national terminology databases of the EU member countries. This creates potential discrepancies between the terminology used in national institutions and EU bodies and hinders terminology harmonization.

IATE currently does not provide efficient mechanisms to consolidate terms originating from different sources into unified multilingual entries. The consolidation process currently is being implemented manually by terminologists from IATE institutions which takes a great deal of time and is very expensive.

2.5.2 CONSOLIDATION OF STANDARDIZED TERMINOLOGY

Important steps towards an interoperable model of terminology management within an international organization are taking place in ISO (Weissinger, 2007). ISO (International Organization for Standardization) is the world's largest developer and publisher of international standards. ISO is a network of the national standards institutes of 162 countries, one member per country.

The traditional standardization process has been based mainly on the production of standards in the form of documents. An examination of standards development activities shows that their development takes place within technical committees, of which the majority is organized vertically based on industry segment and working topic. Many technical committees have their own subcommittees for terminology. ISO technical committees and subcommittees develop and use their own databases to support their work. National members of ISO have started their own initiatives to offer collections of national and international standards terminology (e.g. DIN-Term of the German Institute for Standardization).

Besides terms ISO also deals with other kinds of standardized concept designators: standardized code sets like country codes (ISO 3166), language codes (ISO 639) or currency codes (ISO 4217). These code sets are published in the form of standards documents. The following standardized items are part of ISO standardization activities:

- Terms and definitions;
- Graphical symbols;
- Codes (language, country, currency etc.);
- Units of measurement;
- Product properties;
- Data dictionaries of various types, and others.

It can be seen that a large part of ISO activities are related to the standardization of concepts and their designators. Concept may be designated by different forms of representations as linguistic or verbal units (i.e. terms), graphical symbols/icons and codes.

ISO terms and standardized codes are highly dispersed among numerous standards and proprietary databases resulting in substantial redundancy in investigating and defining terms and definitions, the lack of a single approach makes it very hard for both volunteers developing standards and the industry using standards to make efficient use of existing databases and avoid misinterpretation. Different registration and licensing schemes, inconsistent usage and unclear intellectual property rights statements often confuse the potential users of existing databases (Pohn & Weissinger, 2008).

To solve this problem in 2007 ISO decided to start development of an ISO Concept database (ISO/CDB). It was first released for public use at the end of 2009. The concept database is comprised of content from existing ISO standards, but is also intended to provide a platform for the development of new standards as well as the maintenance of existing standards. It contains not only terms and definitions but also graphical symbols and codes.

The main functions and expectations for the ISO/CDB are to consolidate and host information related to standardized concepts, i.e. itemized standardization units which are part of ISO standards or currently subject to standardization. Development of ISO/CDB follows the broader idea of standards in the form of a database of standardized concept-based items.

In the current version ISO/CDB supports read-only access to terms and definitions, graphical symbols and codes (country, currency, language and script). Terms in the ISO/CDB are taken from terminological entries in standards, which are usually contained in a specific clause with a header like "Terms and Definitions" or similar. With regard to terminological data categories, the ISO/CDB is based on the standard ISO 10241 *Terminological entries in standards*. The ISO/CDB allows to search for concepts in more than 18 000 ISO standards.

End users can access ISO/CDB through an online public information layer, data download, and API-access. Part of the data will be available for free, for other data a subscription mechanism id is planned with webstore-integration.

Through the online public information layer users can get the following information about the searched terms:

- The number of items (search results) found;
- The unique ID of this entry in the ISO/CDB;
- The current stage of the standard in which the entry occurs;
- The term;
- The reference number of the standard in which the term appears;
- The entry number of the terminological entry in the standard;
- The definition of the term in this standard.

Additional information about the standard is provided in a pop-up window when the user moves the cursor over the standards:

- Reference number;
- The title of the standard in English;
- The title of the standard in French;
- The number of the edition of this standard;
- The current stage of the standard in which the entry occurs;
- The committee responsible for the standard;
- The ICS-class(-es) assigned to the standard.

Currently terms in the ISO/CDB are in English only[IS7] and extension to other languages is planned to extend the content to other languages in the future.

The ability to search and compare concepts from the complete collection of ISO standards is expected to help ISO committees in the development and maintenance of their standards and contribute to increasing the consistency among standards. Through ISO/CDB experts are able to access and compare all the instances of terms and definitions used by all ISO committees (either as a single word, or as a combination of words). This enhances access to, and the sharing of, knowledge. This in turn helps to improve content quality, and prevent duplications or inconsistencies.

We can see that ISO/CDB serves as an example of a successful international terminology consolidation activity. It clearly demonstrates the possibility to consolidate terms from a large number of sources from multiple domains. The novel aspect is the inclusion in ISO/CDB not only of terminological information but also other kinds of standardized concept designators like standardized codes and symbols.

At the same time the possibility to generalize this approach to other terminological sources and institutional environments is limited. ISO consolidates terms in a highly regulated environment of international standardization. The structure of terminology standards and term sections in standards are precisely defined and strictly followed (ISO 10241, 1992). This ensures unified logical data structure and homogeneity of terminology sources that greatly facilitates consolidation work. ISO/CDB has not yet solved the task of multilingual terminology consolidation and provides English terms only. It is mostly oriented to standardization tasks and does not provide mechanisms to facilitate terminology access for translation and documentation work.

2.5.3 CONSOLIDATION OF MULTILINGUAL LEGAL TERMINOLOGY

Legal terminology poses a challenging problem for multilingual terminology consolidation as terms are bound to different legal systems and in many cases do not share a common meaning. The LexALP project (Lyding et al., 2006; Chiocchetti & Voltmer, 2008) is focused on the consolidation and harmonization of legal terminology used by the Alpine Convention.

The Alpine Convention is an international treaty signed by all states of the Alpine territory (France, Monaco, Switzerland Liechtenstein, Austria, Germany, Italy and Slovenia) for the protection of landscape and sustainable development of this mountain area. The member states speak four different languages, namely French, German, Italian, and Slovene and have different legal systems and traditions. To ensure effective implementation and communication of the Alpine Convention there is the need for a systematization, consolidation and harmonization of terminology and clear translation equivalence in all four languages.

LexALP term bank[2] includes manually revised, elaborated and validated (harmonised) quadrilingual information on the legal terminology extracted from legal documents of the Alpine Convention countries.

LexALP separates terms from different legal systems in so called 'volumes' as in most cases terms with the same denomination (same lexical form) have differences in their meaning rooted in the respective legal systems. In such a way multilingual terms from the Swiss legal system are kept in separate volumes from lexically equivalent French and German terms.

Terms with different denominations but conveying the same 'meaning' (concept) are also represented using different entries. In this way, synonyms, short forms, abbreviations etc. are stored in separate terminological entries and, if necessary, linked to the relative entries with the full terms through a synonymy relation. However, users have no direct access to these linked data, this must be done via the search interface.

The data categories present in the term bank are denomination/term, definition, context, note, sources, and grammatical information to the term, harmonization status, processing status, geographical usage, frequency and domain, according to a proprietary domain classification structure. Terminologists are given the possibility of writing general comments to the entry. Each term is created in its 'volume' and all available information is provided.

As soon as one or all equivalents in the other languages are available too, a terminologist can create links between these terms. These relations may lead to simple strings of texts (as in the given example) or to autonomous term entries in the dictionary by the use of the `termref` attribute. Simple relations are used, for example, to include information about rejected synonym forms of the term (Sérasset et al., 2006).

For establishing direct translation relation between harmonized equivalent term entries, `termref` relation is used. This is used mostly to link harmonized

[2] http://lexalp.eurac.edu

multilingual terms from the Alpine Convention, which is considered as a legal system expressed in four languages. A different relation `axieref` is used to establish non-direct interlingual equivalence between term entries. It is used to indirectly link national legal terms to the Alpine Convention.

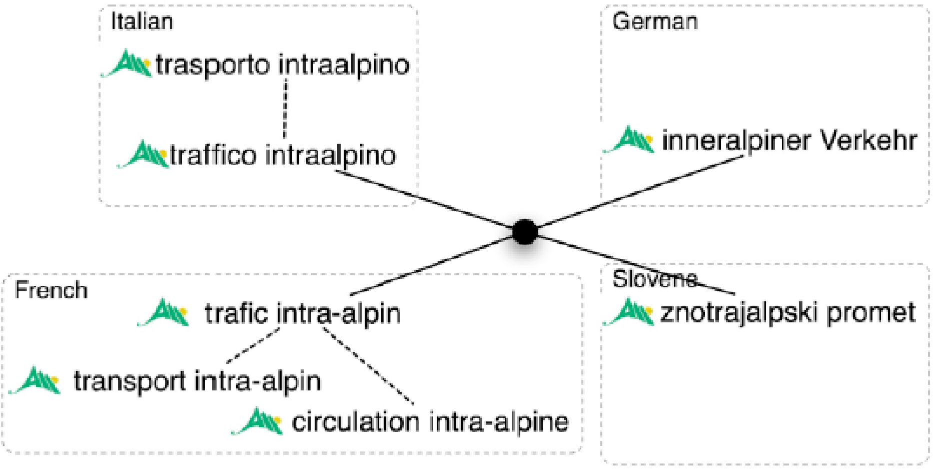


Figure 2 Example of LexALP set of Alpine Convention terms and their relations.

Figure 2 shows an example of direct translation of harmonized terms in four languages marked with a solid line. Corresponding XML representation is showed in Figure 4.

With dashed lines in Figure 2 simple relations to rejected forms are displayed for the Italian term *transporto intraalpino* and French terms *transport intra-alpin* and *circulation intra-alpine*. Figure 3 shows the corresponding XML representation for French term *trafic intra-alpin.*

```
<entry id="fra.trafic_intra-alpin.1010743.e"
             lang="fra"
             legalSystem="AC"
             process_status="FINALISED"
             status="HARMONISED">
       <term>trafic intra-alpin</term>
       <grammar>n.m.</grammar>
       <domain>Transport</domain>
       <usage frequency="common"
             geographical-code="INT"
             technical="false"/>
       <relatedTerm isHarmonised="false"
             relationToTerm="Synonym"
             termref="">
             transport intra-alpin
       </relatedTerm>
       <relatedTerm isHarmonised="false"
             relationToTerm="Synonym"
             termref="">
             circulation intra-alpine
       </relatedTerm>
       <definition>
             [T]rafic constitue de trajets ayant leur
             point de depart et/ou darrivee a
             l’interieurde l’espace alpin.
       </definition>
       <source url="">Prot. Transp., art. 2</source>
       <context url="http://www...">
             Des projets routiers `a grand d´ebit pour
             le trafic intra-alpin peuvent ˆetre r´ealis´es,
             si [...].
       </context>
</entry>
```

Figure 3 LexALP XML form of the term *trafic intra-alpin.*

```
<axie id="axi..1011424.e">
       <termref
             idref="ita.traffico_intraalpino.1010654.e"
             lang="ita"/>
       <termref
             idref="fra.trafic_intra-alpin.1010743.e"
             lang="fra"/>
       <termref
             idref="deu.inneralpiner_Verkehr.1011065.e"
             lang="deu"/>
       <termref
             idref="slo.znotrajalpski_promet.1011132.e"
             lang="slo"/>
       <axieref idref=""/>
       <misc></misc>
</axie>
```

Figure 4 LexALP XML representation of the term relationships from Figure 2.

Besides the terminology database, the LexALP information system also includes a multilingual corpus and the bibliographic database. The corpus provides a reference source to retrieve contextual examples of term usage in legal documents. The bibliographic database includes full information on the text excerpts cited in the term bank and the metadata on corpus documents.

LexALP has successfully consolidated and harmonized terms in a relatively narrow area – terms from the Alpine Convention legal documents in subject fields *Spatial Planning* and *Sustainable Development*. About 500 concepts were harmonized covering about 2000 terms. The LexALP approach involves a centralized system with an extensive manual process in term entry preparation, and harmonization.

In our view this approach is applicable only in narrow domains with relatively small number of terms and an established cooperation structure between participating parties.

2.6 TERMINOLOGY CONSOLIDATION IN NATIONAL DATA BANKS

A national terminology database (or term bank) can contain monolingual or multilingual terminological data and can be established at country, language community or local levels depending on the needs of the respective communities. In terminology planning and in particular in the framework of a national terminology policy, a national terminology database often is used as one of the primary tools for the implementation of that policy (Infoterm, 2005).

(McNaught, 1993) proposes the following characteristics for the national term bank: multifunctional, multilingual, multidisciplinary and widely accessible. He stresses the importance of tight user involvement to ensure that terminology is acquired, elaborated and disseminated in fields and in languages of immediate relevance to users, that services provided are relevant to user's needs and are as user-friendly as possible.

An example of the important role national term bank plays in language policy is the Law on the Term Bank of the Republic of Lithuania adapted by the Lithuanian Parliament (Seimas, 2003). The purpose of the Term Bank is to ensure a consistent usage of normalized Lithuanian terms, especially in the legislative documents of the Republic of Lithuania, to create a common informational system for various state institutions, and to provide access to national terminology data to anybody who is interested in it, including the option to provide data.

Some examples of national terminology banks are given in Table 1.

Country	Term bank	Institution
Finland	TEPA	The Finnish Terminology Centre TSK
Sweden	Rikstermbanken	Swedish national center for terminology TNC
Poland	PolTerm	Polish translators society TEPIS
Latvia	Termnet.lv	Terminology Commission of Academy of Sciences of Latvia
	AkadTerm (termini.lza.lv/akadterm)	Terminology Commission of Academy of Sciences of Latvia
Lithuania	National Terminology Bank	State Language Commission of Lithuania

Table 1 Examples of national term banks.

To characterize the problems and solutions in consolidating national terminology, we will provide a detailed overview of the development of the Latvian national term bank Termnet.lv based on (Vasiļjevs & Skadiņš, 2004; Vasiļjevs & Rirdance, 2008) in the following chapter.

2.6.1 TERMINOLOGY CONSOLIDATION IN LATVIA

Latvian terminology has encountered significant changes in last decades. During the Soviet period Latvian terminology development was to a large extent determined by the Russian language and by new terms requirements of the Soviet political and economic system. After Latvia regained independence and Latvian was established as the official language in Latvia, new requirements and challenges were faced for the development of national terminology. It was recognized early that the Latvian language should serve as a precise instrument for work and communications in all fields. For this purpose, unambiguous, harmonious, and widely accepted terminology incorporating the large number of new concepts rapidly appearing in today’s world is required.

The Terminology Commission of the Academy of Sciences of Latvia (LZA TK) is the official institution responsible for the development of Latvian terminology. Currently there are 23 subcommittees that cover a large spectrum of fields from botany to information and communication technologies. Translation and Terminology Centre (TTC)[3], the official institution for EU related translations, was also active in terminology development. TTC prepared proposals for a unified terminology appropriate to the Latvian language, for use in translation of EU legislation, and NATO documents into Latvian.

In the mid-1990-ties it was realized in Latvia that a public web database is the most effective way to make new terms available to everybody who needs them. The first online terminology database was provided by TTC[4]. It includes a large number of official terms collected from different printed publications and digitalized. This extensive work was accomplished by the Laboratory of Artificial Intelligence at the Institute of Mathematics and Informatics of the University of Latvia in the late 1990-ties.

Soon after this database was developed, a number of terminology sites were opened dedicated to a particular field. Not surprisingly, the IT community was the most active in this respect. The Riga Information Technology Institute – the host of the IT&T terminology sub-commission – was the first to publish proposed and accepted terms on their web site. A technologically advanced online IT&T terminology database was developed and maintained by Eduards Cauna[5]. The IT company Tilde developed both a separate online terminology database and integrated terminology data in an online translation dictionary.

The success of the TTC database and the positive experience of the use of new technologies in IT terminology development led to the idea of creating an integrated online system reflecting the terminology process in Latvia. This idea was initiated by the chairman of the IT&T Terminology subcommission, then president of Latvian Information Technology and Telecommunication Association (LITTA), Prof. Juris Borzovs.

[3] In 2009 TTC was reorganized and merged with the Center of Official Language (*Valsts valodas centrs*) which took over functions of TTC.
[4] http://completedb.ttc.lv
[5] http://www.termini.lv

Some of the goals that were set for this online terminology system were the following:

- **Representation for all fields** to provide one source for all official terms accepted by the Terminology Commission;
- **Instant updates** to publish new terms immediately after their approval;
- **Free access** to official terms, authorized access to internal workspace of terminology institutions;
- **Terminology development workflow.** There are many people from Terminology Commission, TTC, academic institutions, government sector and private business who are involved in development of new terminology. Online technologies can serve as a tool to check consistency of terminology across different fields, to discuss different suggestions, to organize workflow of preparing a new term from an initial suggestion through discussions and proposals to officially accepted term;
- **Offline databases** for subcommissions to enable maintaining separate work databases for particular subcommissions;
- **Data exchange** with offline databases and import/export options to different formats.

The initiative for the terminology portal was launched by LITTA in 2003. The portal was developed and is currently maintained by the aforementioned company Tilde.

The system in use is an integrated platform where terms are developed and managed by LZA TK in the Trados MultiTerm [IS8]environment and published by Tilde in its terminology portal. Trados MultiTerm export format is suggested[IS9] as the data exchange format. This platform architecture solves many issues in terminology development. Each TZA TK subcommittee owns its terminology database; they can use feature-rich, powerful and industry-standard software to manage local databases. Each subcommittee can store unapproved terms, internal discussions and other non-public information in these local databases. These databases can even be exchanged with their partners.

When terms are ready for publishing they are exported to Trados MultiTerm export format and sent to the Portal Moderator who publishes them on the portal. Apart from the possibility of publishing approved terms, the terminology portal can also be used as a discussion, announcement, and document exchange site. Hence it can be used to exchange information and opinions among LZA TK members, subcommittees and terminology users.

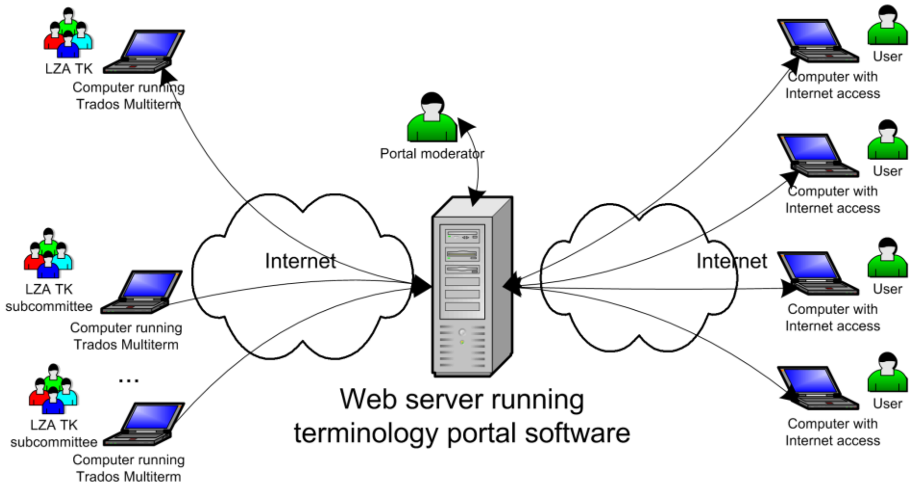

LZA TK
Computer running Trados Multiterm
LZA TK subcommittee
Computer running Trados Multiterm
...
LZA TK subcommittee
Computer running Trados Multiterm
Internet
Portal moderator
Web server running terminology portal software
Internet
User
Computer with Internet access
User
Computer with Internet access
User
Computer with Internet access
User
Computer with Internet access

Figure 5 Main user groups of the Latvian terminology portal Termnet.lv

Although only the portal administrator (moderator) can add new terms to the database, inclusion of new terms can (or rather must) be initiated by the subcommittee which develops them. When a certain amount of new terms is approved and ready for publishing, the subcommittee sends them to portal administrator. There are several reasons why only the administrator should add new terms. Term insertion is related to several issues: old databases must be backed up, server performance can be lost while terms are being inserted, the database is not searchable during insertion. The administrator knows all the procedures including the best time to make updates so that portal users would be affected as little as possible.

Appropriate support activities must be performed to ensure continuous work of portal. Support means technical support for users: both members of LZA TK and all users who just query the web database. It also means maintaining the hardware and software of the web server, making regular backups, moderating the discussions, etc.

Cooperating with the Translation and Terminology Center (TTC), Tilde has added around 115 000 terms from more than 22 fields to the portal.

All data accumulated in the portal (terms, documents, discussions etc.) is stored in a Microsoft SQL Server database. Software is developed using ASP.NET and based on industry standards such as XML/XSL.

The portal is designed for easy administrator customization. It provides different options for different user groups. There is information and features accessible to the public, as well as ones available only for authorized users. The public part of the portal contains termbases and provides search facilities. It also contains various public notices and documents (such as the protocols of LZA TK meetings) in read-only mode. General discussions on terms and terminology are also public. Authorized users can access the non-public parts of the portal available to specific user groups. The administrator can define access rights for each user group for each section of the portal. For example, if the user is a member of a LZA TK subcommittee, he/she can access the document library of the particular subcommittee (editing mode), subcommittee proprietary discussions, notice boards etc.

Users can query the term database in many ways. They can query the full database containing terms from all fields, and they can choose a field and search only the particular database. There are currently around 145 000 terms covering 35 different fields in the database. Users can search in any language presented in the portal and they can even search terms in all languages.

Support of inflectional forms while searching is very important for the Latvian language. The portal has integrated Latvian morphology, which enables the user to find terms even if they are not in the base form. It is important because terms in the definitions are usually not in the base form as Latvian is a highly inflected language.

The portal supports hyperlinks and pictures in the definitions of terms. It enables the terminologists to make descriptions richer and to show relations between terms.

The terminology portal is closely integrated with the reference portal letonika.lv. Letonika.lv contains general usage translating dictionaries (Latvian↔English, Latvian↔Russian, Latvian↔German) and encyclopedias, providing the users with additional information.

The implemented system provides an effective and powerful way to significantly improve the process of terminology development in Latvia.

2.6.2 TERMINOLOGY CONSOLIDATION IN CANADA

TERMIUM, sponsored by the government of Canada, is one of the world’s largest term banks. TERMIUM was established in the early 1970-ties with three basic objectives: to collect terminological documentation, which already exist in various forms, to promote a methodical approach to terminology research in different sectors, and to make all information available to clients through fast and efficient processing. Already in the early 70-ties the creators of the database understood the potential of the computerized terminology system to address such problems as lack of coordination in terminology research, inaccessibility of terminological information, the burden of terminological research during the translation process which takes at least a third of a translator's time, slow dissemination of terminology through printed media (Dubuc, 1972).

Taking into account an early advancement and high general awareness of the role of terminology in bilingual Canada, it is no surprise that nowadays TERMIUM (including its latest version TERMIUM Plus) is among the largest and the most used terminology databases in the world. It is maintained continuously with approximately 50 000 modifications per year such as the creation of new records, deletion of outdated data and expansion of existing records.

TERMIUM includes 3 900 000 terms with definitions, contexts, examples of usage, observations and phraseological units. Most of the terms are in English and French but there are also about 100 000 terms in Spanish. It covers a very diverse spectrum of subject fields so the term bank creators claim that "almost every field of human endeavour is covered"[6].

Among other fields TERMIUM includes standardized English and French terminology from different national and international standards.

TERMIUM provides the following information related to terms: synonyms, acronyms, abbreviations, definitions, contexts, phraseology units, examples of usage and observations. TERMIUM groups definitions, explanations and contexts together as descriptive information in the textual support data category.

For term bank users the following information is provided:

- the subject field to which the concept belongs;
- the languages dealt with: English, French or Spanish;

[6] http://www.btb.termiumplus.gc.ca/site/termium.php?lang=eng&cont=005 (accessed 20.06.2010)

- the terms, for example: *'terminology record,' 'fiche de terminologie,' and 'fiche terminologique';*
- the term usage labels, for example: '*officially approved', ' feminine';*
- the textual supports, for example: *'A medium for recording terminological data', 'Support sur lequel sont consignées selon un protocole établi les données terminologiques relatives à une notion'.*

The content of the database is accessible to translators, technical writers and other professionals. Several spin-off products are also developed, such as an on-line linguistic tool *TERMIUM Plus*® which is built on top of the term bank, providing writing assistance facilities in English and French and giving access to 13 electronic language resources.

2.7 TERMINOLOGY CONSOLIDATION AT LOCAL AND MULTINATIONAL ORGANIZATIONS

Local terminology work is performed by organizations such as translation agencies, research institutes, local companies, etc.

2.7.1 TEMINOLOGY CONSOLIDATION AT MULTINATIONAL COMPANIES

Terminology work is an important part of the activities at large multinational companies to ensure consistent usage of terms in product documentation, interface, and communications with customers in target language markets.

Global IT and software companies are among the most advanced in implementing terminology consolidation, distribution and exchange mechanisms. Companies like IBM, Microsoft, Oracle, Sun, Apple, Novell and others have established in-house terminology management systems.

Let us provide a few examples of terminology work in SAP and Microsoft based on (Rirdance & Vasiljevs, 2006).

SAP has a one-stop terminology interface integrated into its SAP R/3 system (Transaction SE63 - Translation Environment). Similarly to other SAP products, the terminology tool also offers a wide range of user privilege management features, and terms can have different statuses of approval until they reach the status of approved.

Some SAP translators – especially the localizers – get access to this software and can suggest new terms, others are provided with bilingual MultiTerm glossaries to get help in their work. These glossaries are also available for sale and contain entries in Bulgarian, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Polish, Portuguese, Russian, Slovakian, Slovenian, Spanish, Swedish and Turkish. So far, the database contains about 650,000 terminology entries in 20 European languages, and nearly 16,000 definitions of SAP concepts. SAP has made public some part of these terms with definitions in English and German.

At SAP, the privileged languages are English and German, and all terms need to have equivalents in all languages. Terminological entries are usually created by knowledge brokers and authors of texts as well as in English or German. Entries include not only software-related entries (screen captions, etc.) but also entries appearing in training course materials and marketing materials. Translators can also enter new terms, but superusers – consultants – need to approve them.

Entries include a wide range of information, including definition and part-of-speech information, but the emphasis falls on the source of the term.

At SAP, therefore, all users regardless of their nationality use the same terminology database, there are no competing databases.

Microsoft employs a particular term registration process. Microsoft follows a systematic approach to software-encoded terminology, which starts during development. Developers create terminology during program design and development in an intuitive/metaphoric way. Important terms – e.g. brand names, major technology names - are also reviewed by other personnel, sometimes even tested in a public opinion poll. The language of all source terms is English. The creation of the initial termbase is automatic: their own localization software extracts all the string resources from the products.

In Microsoft definitions are not an integral part of their terminology. They use multilingual terms and usage examples to provide consistent localization for their products.

Microsoft employs a few (1-3) terminologists for every language they provide a product version for. These terminologists create a core termbase for each and every product, building on the terminology of former products and user responses. The core termbase is then sent out to localizers, who have to create local versions of their products - and their terminology.

Most of the terms employed appear in screen captions. Termbases (in the form of source string – target string) are unique for each and every product and product version, and contain the screen captions and some help-specific terms. Non-software-related terms (which are only a few in number) are not collected in a single termbase, but Microsoft Press, the official publishing house of Microsoft, regularly updates its Microsoft Press Dictionary.

Microsoft provides online public access to its terminology database through Microsoft Language Portal[7].

2.7.2 ONTOLOGY-BASED CONSOLIDATION OF MEDICAL TERMINOLOGY

ECDC Core Terminology Server project serves as an early experience in coping with the depth of concept related information.

The central goal of the Core Terminology Server project at the European Centre for Disease Prevention and Control (ECDC CTS) is to support creation, maintenance and dissemination of ECDC terminology and provide terminology services on concepts that are related to the activities of ECDC to both human users and software applications, within ECDC internally as well as to external users and applications. The first version of CTS was released in June 2008.

The European Centre for Disease prevention and Control was established in 2004 by a European Parliament and Council regulation, to identify, asses and communicate current and emerging threats to human health from communicable disease. One of the essential characteristics of ECDC is its interdisciplinary nature. ECDC activities are at the crossroads of different branches of medicine (e.g. public health, communicable diseases, microbiology, lab science) and of legal environments, of policy making of the EU, that of the member states and the World Health Organization (WHO). All these disciplines are practiced in the environment of each and every official language of the European Union. As ECDC defines itself as a pan-European undertaking, the proper ’labeling’, that is, terminology of what it does and how it is communicated

[7] http://www.microsoft.com/language

becomes a problem. For example, wording in EU legal texts is different from wording in WHO International Health Regulations; terminology used in disease surveillance systems is different from terminology used in the EWRS (early warning and response system) and other systems. Obstacles of setting up shared databases and information services lies at the fundamental level: the same concepts have different designations in different contexts. In spite of many terminology systems available and although parts of them are reusable, there is no ready-made terminology at the specific crossroads of the ECDC mission. Therefore ECDC decided to build up its own terminology using all externally available sources, to be able to assist its own staff and partner organizations with a set of terminology services that provide human and software readable explanations, characteristics, features, code sets where available, synonyms, translations and other services to all concepts related to activities of ECDC.

As stated in (Balkanyi, 2007), ECDC CTS is primarily designed for use by ECDC experts, various health data and medical professionals in EU member states, such as public health data administrators, health ICT system developers, epidemiologists, public health experts working with terminologies, and others. ECDC plans to establish and run a service for terminology in ECDC CTS. ECDC CTS also supports machine users that connect to ECDC CTS via a Web service and use its terminology content for other software applications both in-house and later externally as well.

The backbone of the ECDC terminology system is an ontology, or the semantic network, that consists of concepts mapped to the terminology content of ECDC CTS, consisting of application specific sets of terms (called value-sets), represented as categories. All value set categories are mapped to at least one concept in the ECDC CTS ontology. Value sets represent terms of a certain subject field, or domain or terms used by a client application.

An important requirement for the system is the ability to build and trace relations among concepts within ECDC semantic network. This means that, behind the enriched specialized vocabulary, users have an ontology that maps relations among concepts and supports easy navigation in the ”concept-space”, allowing users to go to broader or narrower terms along their hierarchy and follow relations in the net of related concepts. The rationale for an ontology backbone to the systems lies in making full use of description logics support, such as (semi)automatic reasoning and decision support functions expected to be implemented as a follow-up.

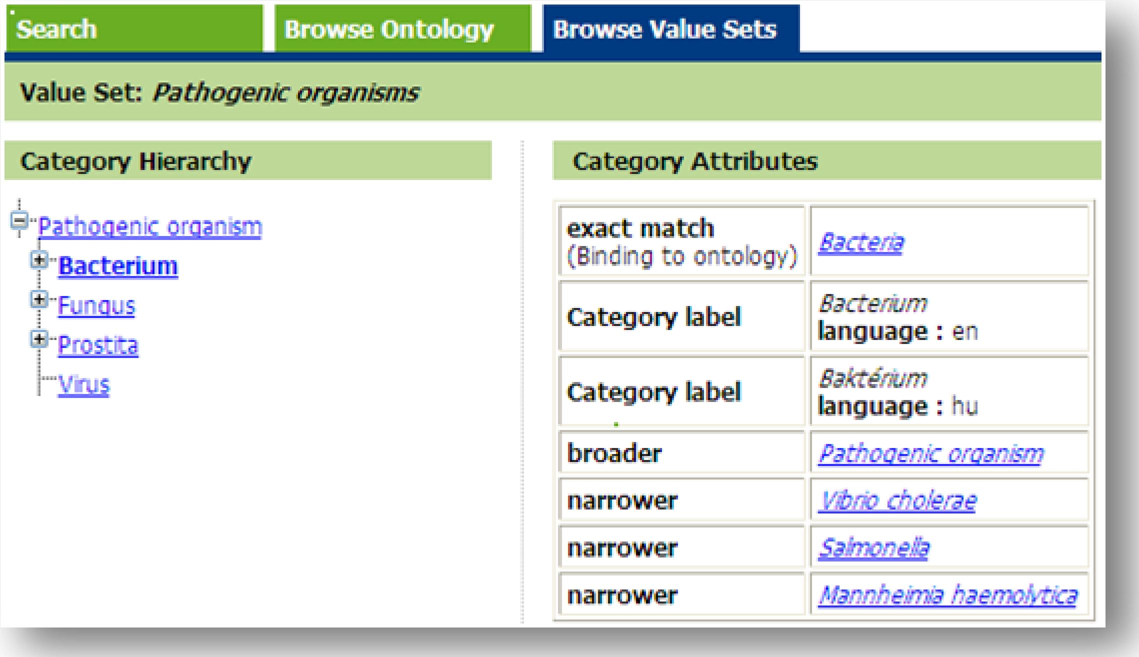
Taken the novelty of ECDC CTS project, it makes use of a number of new and emerging standards. Many of these specifications are based or built atop of the Resource Description Framework (RDF) by World Wide Web Consortium (W3C). The RDF derived a SKOS (Simple Knowledge Organisation System) specification that provides formalism to represent structured concept systems and is widely used throughout the ECDC CTS, for data storage and maintenance as well as import and export of value sets. The core ontology, although created and maintained in OWL, is converted to SKOS before being imported into the system.

The Web Ontology Language (OWL) is used for ontology representation in ECDC CTS. It is designed to represent Web ontologies in a machine-interpretable way and can serve as an exchange format for ontology information exchanged between different systems. However, to maintain data integrity and to ensure conversion to SKOS, ECDC CTS supports limited OWL implementation.

Another format applied as input and output of terminology value sets in ECDC CTS is ClaML (Classification Markup Language). It is based on XML and is adapted as a CEN Technical Specification (TS14463). Although being a relatively new formalism it is already under adoption by institutions such as the World Health Organization for handling disease classifications and is hence highly relevant in the ECDC area of activities.

For terminology work and information management *ISO Guidelines for the establishment and development of multilingual thesauri* (ISO 5964, 1985) and the abstract model for Dublin Core metadata (DCMI, 2007) are employed.

The ECDC CTS provides information about terminology systems represented as value sets, while the concepts are represented as categories. Each category is mapped exactly to one concept in the underlying core ontology, which is represented as a semantic network in the ECDC CTS. The conceptual elements (value sets, categories, ontology, concepts and relations) are represented in SKOS. Since some of the terminologies used are coming from WHO classification systems (e.g. ICD10), the system supports export/import of value sets in ClaML format.



Value Set: *Pathogenic organisms*

Category Hierarchy

- Pathogenic organism
  - Bacterium
  - Fungus
  - Prostita
  - Virus

Category Attributes

exact match (Binding to ontology)	*Bacteria*
Category label	*Bacterium* **language :** en
Category label	*Baktérium* **language :** hu
broader	*Pathogenic organism*
narrower	*Vibrio cholerae*
narrower	*Salmonella*
narrower	*Mannheimia haemolytica*

Figure 6 Example of the hierarchical and attributes views in ECDC CTS

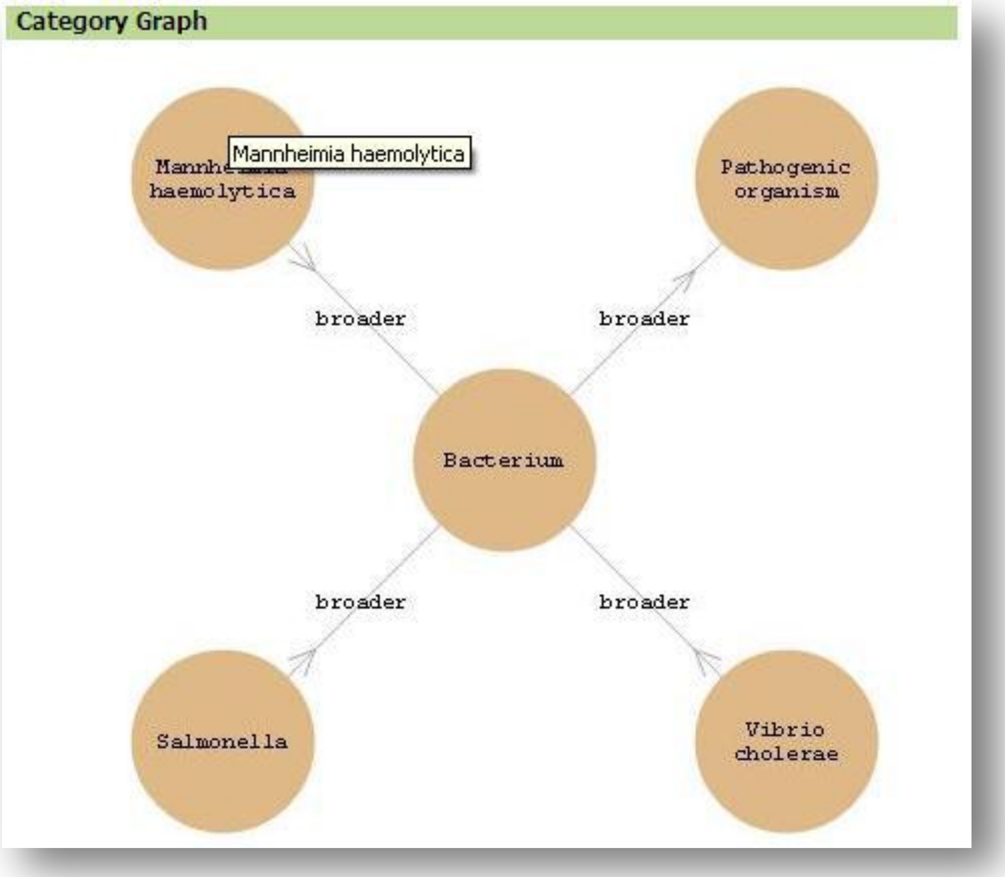


Figure 7 ECDC CTS Semantic relationships in graph view

ECDC CTS provides ample opportunities for the user to view information related to a certain category, or a term:

- the tree view displays hierarchical information, showing the place of a concept in its hierarchy;
- the attributes section provides full information about it, such as its natural language labels, Dublin Core attributes, ontology binding, and others; the graph view shows the semantic neighborhood of the concept in a graphical form.

Examples of these two views are shown in Figure 6 and Figure 7.

In the current phase of the ECDC CTS project, the terminology resources to be included are predominantly in the English language only. However, it is possible to include as many languages as needed by adding additional concept labels for each language, as shown in Figure 6, where both the English and the Hungarian terms are displayed.

Although current activities in terminology consolidation are mostly focused on unification and harmonization of dispersed heterogeneous terminology resources, ECDC CTS development is an example of an emerging need to enrich traditional terminology with machine-readable semantic information. This will provide ample opportunities for different semantic applications well beyond the traditional use of terminology systems. However, ontological enrichment and implementation of semantic representation specifications pose serious challenges in the evolution of traditional terminology databases.

2.8 ISO STANDARDS IN THE TERMINOLOGY FIELD

Standardization is essential for consolidation of diverse terminology resources and ensuring exchangeability of terminology data. During the EuroTermBank project standards for terminology data processing were assessed and applied for data modeling and data interchange interfaces.

The most recognized standardization body is Technical Committee 37 of International Standardization Organization (ISO TC37) *Terminology and other language and* content resources. The scope of this ISO technical committee is "the standardization of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity" (Warburton, 2007). It consists of four subcommittees:

1. Principles and methods
2. Terminographical and lexicographical working methods
3. Systems to manage terminology, knowledge and content
4. Language resource management

A number of standards developed by ISO TC37 describe basic principles for terminology data modeling, processing, storage and interchange.

For purposes of storage and retrieval, terminology data is organized into terminological entries. Each entry includes information related to the single concept. This concept-oriented approach differs from widespread practice in many dictionaries to organize entries around lexical units. To consolidated terminology entries in different languages and from different sources it is necessary to group them around abstract language independent concepts.

Individual terminological entries consist of data items according to a chosen data model and data category. The International Standard ISO 12620 *Computer applications in terminology – Data categories* specifies data categories for recording terminological information in both computerized and non-computerized environments and for the interchange and retrieval of terminological information independent of the local software applications or hardware environments in which these data categories are used. The use of uniform standard-compliant data category names and definitions greatly facilitates interchange of data between different systems and enhances the reusability of data.

For the interchange of terminological data an international standard ISO 12200 *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF)* has been developed. It allows the distinct identification of separate data sets and data categories as well as their dependencies and relations. The format relies heavily on the data category names and definitions contained in the standard ISO 12620. MARTIF is based on ISO 8879 *Standard Generalized Markup Language* (SGML).

MARTIF provides an open, flexible mechanism for exchanging data between different terminology management systems. The main body of the MARTIF standard specifies the formalism to be used in preparing terminology data collections for interchange by defining the SGML Document Type Definition (DTD) and listing the appropriate tags (markup) used to structure the data. Normative Standard also specifies the markup for the individual terminological data categories to be used in the MARTIF environment, based on ISO 12620.

International standard ISO 16642 *Computer applications in terminology - Terminological markup framework (TMF)* facilitates the use and re-use of terminological data collections, taking into account the real-live conditions of different formats, database environments and term-bank systems as well as the various data models the collections are based on. The standard also addresses the need to provide better connections between terminological databases and other lexical resources used, for instance, in machine translation or natural language processing.

Localization Industry Standards Association (LISA) has developed an industry standard TBX (short for TermBase eXchange). It is a very practical terminology exchange format that is compliant with the terminology markup framework TMF. TBX is based on the TMF structural meta model; it specifies a set of data categories from ISO 12620 and adopts an XML style compatible with MARTIF.

3. METHODOLOGY FOR TERMINOLOGY CONSOLIDATION

In our research we focus on the problem of consolidation of heterogeneous multilingual terminology resources.

There are a large number of different resources of terminology data in European countries and beyond. At the same time, the overall situation in the global terminology area is characterized by many gaps and problems. Resources are fragmented, located in different institutions and in different format. Much of terminology data is still available only in the form of printed dictionaries and bulletins or stored in card files. In many countries there is a lack of coordination and unified methodology between institutions involved in terminology development leading to inconsistency and poor quality of terminology data, insufficient mechanisms for dissemination of new terminology.

The heterogeneous nature of terminology resources is characterized by differences in data structure, language coverage, organization principles, formatting, storage formats and geographical location.

Although terminology management systems are widely used in practical localization work, it is common to see industry termbases that contain only the source and target term, and perhaps a comment if the source term has multiple possible translations depending on the context (Somers, 2003).

The integrated approach proposed in this research encompasses all the major aspects related to the consolidation problem:

- Requirements analysis in multinational multi-actor and multiuser environment;
- Data modeling principles for terminological information;
- Data storage and data exchange mechanisms;
- Consolidation approach for independently maintained terminology databases;
- Unified representation of dispersed heterogeneous terminology data.

3.1 DEFINING TERMINOLOGY USERS AND THEIR REQUIREMENTS

According to the principles of user-centered design (Beyer & Holtzblatt, 1998) development of a user-oriented information system should be based on extensive analysis of needs, preferences, and limitations of end users.

(Cabré, 1999) recommends to start development of a terminology system with identification of the main user groups and an analysis of their needs. A user survey is recommended for this task:

- Identification of target users;
- Delimitation of the needs of each user group;
- Comparison, coordination, prioritization of the needs identified.

In the framework of our research and related EuroTermBank project, a survey and individual interviews of different target groups of potential users were carried out in 2005, helping to identify user groups and the typical use cases of terminology resources (EuroTermBank Consortium, 2005; Henriksen et al, 2005).

In total 51 questionnaires were completed providing an overview of possible usage cases and corresponding user requirements for particular cases. The most typical usage of online terminological resources is *translation*. These users require single access to multiple data sources and a convenient user interface. Access to terminology databases through integration with popular CAT (computer-assisted translation) tools was also requested.

Another popular usage is *general research* („look up terms”). In this scenario instant access to terminology reference is required during reading and research providing comprehensive information about a terminology entry like subject field, definition, status, target-language equivalents, abbreviations, usage examples etc.

More specific user groups are lexicographers and terminologists. Besides comprehensive information they require advanced filtering and data export features, they are interested in multiple languages and also in online collaboration facilities.

Interviewees were also asked to name the data categories they are usually looking for when exploring terminology resources. In decreasing order of popularity answers included target-language equivalents, definitions, status/authority/authenticity of entries, subject field, usage examples and synonyms. The requirements listed by users also included abbreviations, closely associated terms, source of term, usage notes and other data fields.

Analyzing the results of the survey, it was determined that the system needs to support two main types of users:

- Human users: such as translators, terminologists, lexicographers, various experts and general purpose users
- Machine users: other systems that will connect to EuroTermBank, such as interlinked external terminology databases, CAT tool plug-ins and other web or desktop applications

The human users were further subdivided according to their roles and relationship with the system into the following groups:

- Anonymous User
- Registered Subscriber
- Editor
- Administrator

The survey also identified the necessity to provide a customizable user interface to address the needs of different target groups. Users should be able to choose between a simpler and easier-to-use interface limiting represented data to few languages and data categories, and a comprehensive interface uncovering the richness and complexity of data stored in the system.

Another more recent survey – TTC survey[8] – identified the latest patterns in terminology usage (Gornostay et al., 2010). The survey was widely distributed through targeted mails, mailing lists, forums and other web-channels to different target groups of terminologists, translators, technical writers and other so called language workers.

Responses from 93 participants have been analyzed in this survey: freelance translators (32.7%), researchers (13.3%), staff translators (9.3%), translation project managers (8.7%), editors (8%), terminologists (6%), technical writers (2.7%), translation volunteers (1.3%) and others (18%). The respondents in the “other

[8] Survey was carried out in the framework of EU FP7 project TTC.

category" include language service consultants and analysts, software developers, university lecturers and language teachers, company managers and owners. The geographic coverage of the survey is quite representative – 31 countries all over the world. 27.6% of respondents indicated that they work in technical translation in comparison with 21.3% in software localization, 14.5% in legal translation, 9.5% in mass media translation, 5% in technical writing, 4.5% in literary translation and game localization. 13.1% of respondents work in other translation sectors including medical, business/finance, ontology, aviation, and others.

The majority of respondents usually spend 20-30% of their time working with terminology (60.3%) performing a range of activities with the terminology research (20.7%). Of those respondents, majority (62.9%) perform bilingual terminology management, 23.6% perform multilingual, and 17.6% – monolingual terminology management.

64.5% of respondents use online terminology databases. The most popular linguistic resources for researching terminology are online resources (35%). The top three are: IATE (20.4%), EuroTermBank (19.1%), Microsoft Language Portal (16%).

76.1% of respondents are interested in storing and working with their terminology in an online terminology database and 36.9% of respondents do not mind sharing terms with the community. When using online terminology resources respondents indicate that the following top five features and aspects are important for them:

- Lookup speed (25%);
- Good coverage of terms across languages and domains (20.9%);
- Number of lookups returned as precise as possible (20.9%);
- Hyperlinks (terms can be reverse-looked up by a simple click) (16.3%);
- Saving terms / search history (9.2%).

Respondents also noted that they are interested in "expert forums and discussion groups on the Internet", "domain specific corpora", "consulting terminologists, domain experts, other translators".

In response to the question about what type of terminological information is usually researched the following answers have been received:

- Lexical, translation equivalents, definition in a source language - 22.4%;
- Grammatical: part of speech, inflection etc. – 10.6%;

- Contextual: example sentences etc. – 19.2%;
- Usage: style, usage note, frequency etc. – 15.3%;
- Categorical: subject fields, domains, products etc. – 13.2%;
- Administrative: status, date, author, source etc. – 5.5%;
- Term relations: synonym, antonym, acronym, related terms etc. – 13.1%;
- Other – 0.7%.

3.2 TERMINOLOGY WORK SCENARIO BASED APPROACH IN REQUIREMENTS ANALYSIS

In this section we introduce terminology work scenarios based approach in terminology consolidation basing on (Henriksen et al., 2006).

3.2.1 GOALS AND CONDITIONS IN TERMINOLOGY WORK

The different terminology consolidation projects and activities described in Chapter 2 demonstrate the best practice in particular settings. International standards provide a strong background for a unified and standardized approach in terminology work. However, standards are very general and describe recommendations in a vacuum disconnected from specific goals and preferences and also disconnected from the set of conditions that apply in a given context of particular terminology consolidation needs.

By conditions we refer to the premises or state of things that cannot (or only with much difficulty) be changed. For example a condition might be that all language professionals of a particular organization do not have access to the internet or to terminology tools. Therefore it is necessary not only to investigate how terminology work is actually carried out in different settings, but also to investigate the conditions and goals of the particular terminology settings.

The goals and conditions identified in the EuroTermBank survey were collected and an assessment of the influence of each was determined by assigning scores. The aim of allocating scores was i.a. to identify sets of goals and conditions that typically co-exist as a first step towards the establishment of a number of fixed scenarios that include best practice descriptions for each terminology task[IS10].

The following goals have been identified as having a profound impact on terminology methodologies:

- High quality in general terms - terminology work is based on sound research principles; consistent, non-ambiguous, broadly accepted etc.;
- Harmonization - in many contexts an inherent part of terminology quality criteria;
- Exchangeability - exchange of data between term resources using standard approved exchange methodology;
- Availability - terminology available to external users;
- Speed and up-to-datedness - speed of terminology work and data that are always up-to-date.

Major factors affecting terminology work that were identified are the following:

- Terminology tools - users may or may not have access tools such as corpus/term extraction tools;
- Type of language professionals - may or may not include terminologists and domain experts;
- Financial situation - satisfactory or unsatisfactory;
- Languages in terminology resource - monolingual, bilingual or multilingual;
- Domain coverage - broad spectrum of subject fields or focused to particular domains;
- Purpose (translation, coordination, regulation) - some organizations have translation as their main focus; others like standardization institutions or national terminology regulatory bodies also have coordinating and regulatory obligations.

3.2.2 TERMINOLOGY WORK SCENARIOS

We propose to distinguish 3 levels of terminology work – local, national and international (Henriksen et al., 2006).

On the **local level**, terminology work serves the needs of a particular organization, such as a company, a translation agency, a documentation centre, a research institute, etc. The local level is usually limited to one or a few closely related domains and is primarily concerned with terminology work originating from translation or the creation of documents. Terminology work at the local level is usually limited in scope and the involved people, therefore terminology consolidation is not a major concern at this level (exceptions are multi-national companies and institutions with terminology work spread around the globe). Although interoperability is not among the top priorities at this level, there is a growing awareness about the potential benefits from integration of locally created terminology into a common terminology infrastructure.

Terminology activities on the **national level** are concerned with the monolingual or bilingual terminology work performed on the level of a specific country. Among the basic tasks of institutions involved in national terminology are national standardization and approval of terms, maintenance of national terminology, and development of integrated terminology systems

In some countries like Latvia and Lithuania, national terminology work also serves the normative purpose defining the “official” terminology for use in legislative documents. In other countries, consolidation and coordination are the major foci at the national level.

Exchangeability and harmonization of terminology resources typically are of high to medium priority at this level, as it may involve a complex structure of actors, compliance to national regulatory management of a national term database and a harmonized multi-branch term system.

The **international level** concerns consolidation and harmonization of terms coined at the national and local levels; it involves coordination and management of multilingual terminology in a well-organized infrastructure. Since consolidation of terminology resources is the cornerstone of terminology work at this level, it not only requires rigorous application of existing standards, but also acts as the driving factor behind improvements and development of new standards and approaches.

Terminology collections at the international level are multilingual, this being a differentiator from other levels that are usually focused on one or a few languages. Terminology work at the international level should optimally include coordination of terminology work between the different countries and institutions involved as well as ensure data interoperability and facilitate terminology harmonization.

International scenario	National scenario	Local scenario
Goals		
High quality in general terms	High quality in general terms	Tight time frames coexist with - and put limitations on requirements for - high quality
Harmonization is high priority	Harmonization is high priority	Harmonization is not a priority
Exchangeability is high priority	Exchangeability is high priority/is sometimes not a priority (recommended as high priority)	Exchangeability is often not a priority (recommended as high priority)
Availability is high priority	Availability is high priority	Availability is not a priority
Conditions		
Access to terminology tools	Access/no access to terminology tools	No access to terminology tools
language professionals represented	All types of language professionals represented	Terminologists often not part of terminology developer team
Adequate financial support	Adequate financial support	Often a tight budget
Multilingual	Mono- or bilingual	Usually bi- or multilingual
Broad domain coverage	Broad domain coverage	Focused domain coverage
Coordination (translation)	Coordination (regulation, translation)	Usually translation

Table 2 Goals and conditions in terminology work scenarios.

3.3 EVALUATION AND SELECTION OF RESOURCES

One of the major tasks in terminology consolidation is identification, description, and classification of a large number of existing printed and electronic terminology resources available in different countries and selection of resources for possible inclusion in the consolidated termbase. In this section approaches for evaluation, selection and description of resources are described.

We propose the following criteria for systematic evaluation of terminology resources, making selection and prioritization for inclusion in the database:

A. Only resources related to Language for Special Purposes, Language for General Purposes resources should not be considered terminological resources;

B. Authority, reputation and expertise of the creating institution or person – whether resource is prepared by a group of experts or by an individual expert, whether specialized lexicographers have been involved etc. The institutions or the authors creating terminology resources can be considered a valuable indication of the quality of a collection. When the institution or the author is known for well-founded terminology work and reputed experts of the respective subject field are involved, there is a good chance that the quality of the terminology collection is appropriate. However, just the fact that an institution or an author is not known so far should not be a sufficient reason to exclude their terminology resources from consideration.

   Data originators listed by degree of authoritativeness are:

   - Legal international or national authority determined by legislation or jurisdiction;
   - Officially authorized harmonization-/standardization body;
   - Institution authorized or recognized as a subject field authority;
   - Formally or informally recognized subject-field authority;
   - Non-authoritative terminology source.

C. Methodological approach – observance of relevant national or international standards, completeness of entries (priority for terms with the most fields populated), existence of internal/external validation mechanisms. Central quality criteria are concept orientation, subject field indications and usage notes, alphabetical indices in all languages, abbreviations and definitions.

D. Availability of the data - to make use of the data, either the terminology resources must be freely accessible or the respective copyright holder should be ready to cooperate and to conclude a copyright agreement with the project consortium.

E. Actuality of the data – topicality, frequency of use, date of input or revision. This criterion is closely connected with the respective subject field. For example, in some subject fields old terminology resources of new EU countries include concepts and terms related to outdated realities that are not of general interest today (e.g. soviet-time concepts).

3.4 TERMINOLOGY RESOURCE DESCRIPTION

Any non-trivial resource consolidation activity faces the need to describe and register data resources. To organize, structure, process and analyze this data it is important to use a common format for resource description.

Metadata for lexical and linguistic resources can be specified in different formats. In corpus work the Dublin Core format is widely used (DCMI, 2007). Selecting generic formalism has an advantage of being directly exchangeable with descriptions of other kind of linguistic information. The disadvantage is that specific guidelines for its usage in terminology work are required. Lack of such guidelines leaves room for different application leading to compatibility problems.

For this reason for terminology consolidation we recommend to use the The Terminology Documentation Interchange Format TeDIF. It is specifically targeted to meta information related to terminological information and is more appropriate for terminology work than a generic formalism like Dublin Core.

TeDIF was developed in the framework of the TDCnet project – European Terminology Documentation Centre Network, co-funded by the European Commission. The TeDIF format was developed with the purpose to establish a common format for bibliographical and factual data related to terminology.

TeDIF provides the means to describe bibliographical data like literature (serials, monographs, articles, journals, theses, etc.) and term collections (printed dictionaries, glossaries, thesauri, classifications, terminology databases, etc.).

These include:

- Bibliographical data:

  - Literature (serials, monographs, articles, journals, theses, etc.);
  - Term collections (printed dictionaries, glossaries, thesauri, classifications, terminology databases, etc.).
- Factual data:
  - Corporate entities (organizations, institutions);
  - Persons (experts);
  - Projects;
  - Terminology management software;
  - Events (conferences, workshops);
  - Teaching and training opportunities.

TeDIF is an SGML-based format (Standard Generalized markup Language, ISO 8879:1986) to describe and exchange data. Since TeDIF is also XML-compatible (Extended markup Language, subset of SGML), it is open to the newest developments in markup languages, the usage of Unicode, and an easier conversion to HTML and other formats.

Specialists with sufficient technical skills can prepare resource descriptions directly in the TeDIF format. Easy to use front-end applications can also be designed that facilitates data entry and provides some automated validation of the entered data.

The applicability of TeDIF was evaluated in EuroTermBank project. It was used to describe terminology resources and to import terminology resource meta-data into EuroTermBank database, as well as to consolidate and ~~analyse~~analyze data.

In the EuroTermBank project a special Excel spreadsheet form was created to provide a very easy way for entering data and avoiding possible mistakes. In order to validate and transform Excel files to the TeDIF format a converter utility was programmed.

Practical application of TeDIF showed a need to make a few modifications to this format:

- Possibility to multiply the fields [IS11]describing the author;
- possibility to multiply the fields [IS12]describing the copyright holder according to the number of persons/organizations;
- Addition of a field to indicate the languages of definitions;
- Addition of field to indicate the languages of context information.

3.5 WORKFLOW OF TERMINOLOGY RESOURCE PROCESSING

Terminological collections are organized in a large variety of different formats. To create a consolidated term bank resources should be transformed to a single unified data exchange format. We propose to use TBX as the standardized format to ensure international interoperability.

The transformation of terminological resources may be represented as a workflow in the lifecycle of the terminological resource from the acquisition of the resource until the resource is imported into the system database. This section describes the resource processing workflow basing on (Liedskalniņš , Vasiļjevs, & Rirdance, 2007).

The schematic representation of the processing workflow including transformation and validation processes is shown in Figure 8. It is further described in this section and subsections.

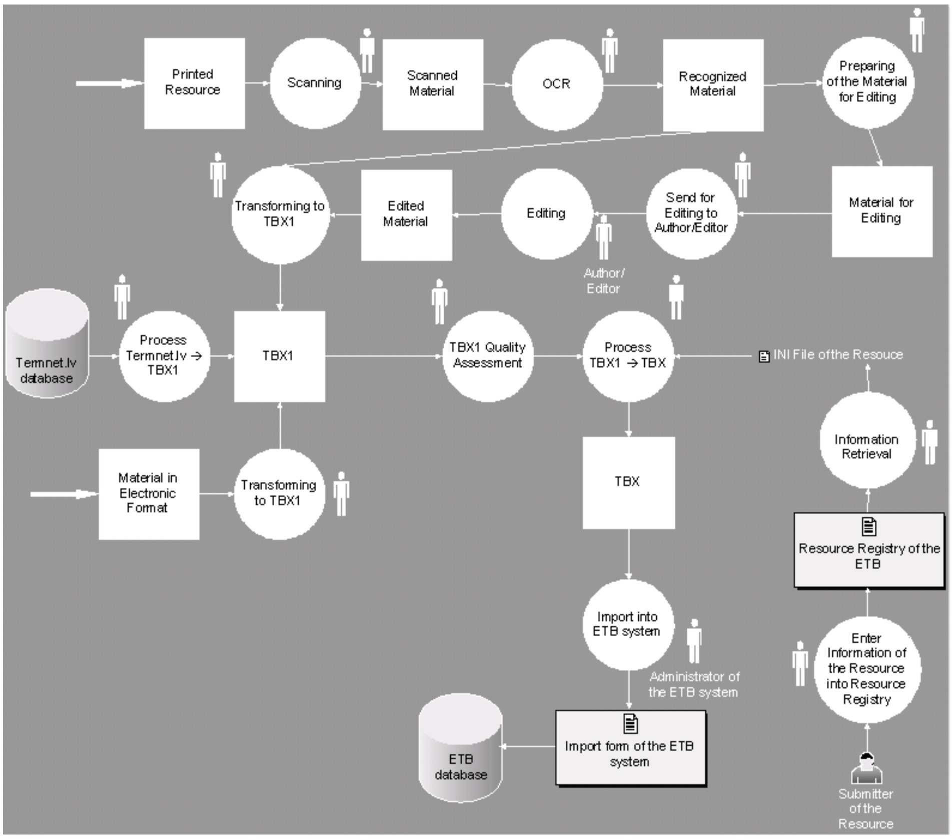


Figure 8. Resource processing workflow.

All resources to be included in this multilingual term bank can be divided into the following four groups: 1) printed resources, 2) individual resources in electronic format, 3) collections of resources gathered from other systems, and 4) resources residing on autonomously maintained termbases.

The processing of printed resources is similar to processing resources in electronic format, except the digitalization step that precedes transformation to TBX.

The last group of resources is not involved in the full resource processing workflow, as they are already processed during data preparation at the host termbase. However, data transformations may be required, unless the resource is in TBX format. The online transformation process of external resources is done by creating a mediator for every external resource. The mediator facilitates the communication process by providing data in TBX format, so that the internal system always gets data in a unified format and does not have to deal with data transformations.

3.5.1 RESOURCE ANALYSIS

The first phase before transformation starts is the analysis of resources. Resources can be evaluated according to the methodology described in Chapter 3.3. The analysis of the data structure is applied to evaluate automation possibilities to transform resource to TBX format. The main characteristics for inferior quality resources (both printed and electronic) that should be avoided are:

- Unformatted/unstructured data;
- No boundaries between different data categories;
- Ambiguous data categories;
- Ambiguous relations between data categories;
- Erroneous/inaccurate data.

Resources that have such characteristics may be included in the term bank, but the possibility of errors must be taken into account and a manual quality check is required. Processing of such resources cannot be done automatically and involves significant manual work.

3.5.2 RESOURCE/DATA TRANSFORMATION

Processing of printed resources starts out with scanning. Optical character recognition (OCR) is used to transform scanned images to text. The OCR process is an important step that determines the resource structure afterwards. Sometimes it is advisable to recognize content as a table, while in other cases it is better to recognize it as plain text. The OCR-ed material typically requires editing. To improve the editing process, the recognized material must be properly prepared for editing, by:

- Identifying misspelled words;
- Highlighting possible errors;
- Explicitly emphasizing the boundaries of data categories.

Direct transformation to TBX format is done after the OCR-ed and prepared material is edited by resource authors or editors. The transformation to the TBX phase is common for both printed and electronic resources. Unfortunately in practice almost every resource has different format and thus must be processed separately.

The processed resources must be validated before they can be imported into system database. The validation process is described in the next section. When a resource fails in the validation process, it is returned to the transformation phase for error correction.

3.5.3 DATA VALIDATION

Initially, validation of a TBX XML document is done using XML schema, to verify that it conforms to TBX format. If this validation fails, there is no need for further validation and the resource must be returned to the transformation phase.

Resources that are formed as valid TBX format XML documents are validated further. In TBX format some of the fields are restricted to a set of predefined values. Transformed resources may be easily validated against these predefined set of values. If such a field contains another value, it is most possibly due to an error and is reported in the validation summary.

As terminological data contains language dependent information, language specific validation rules can be applied. Character validation tests check conformity of textual fields to allowed characters corresponding to particular language.

Spellchecking is the last step in data validation. All words are spellchecked with the spellchecker of each respective language. Output of this validation is a percentage of the words that are not spelled correctly. This information may be analyzed to see if the resource has been processed correctly. If the rate of misspelled word is high[V13], then the boundaries of data categories and terminological entries may be incorrectly determined.

The validation summary is a list of validation results. Some of this information may be examined automatically, but some must be manually revised to determine whether it is an error or not. If the resource completes validation process, it is ready to be imported and stored in the system database.

3.6 TERMINOLOGY DATA STORAGE

There are four main tasks that must be considered when managing hierarchical data – store, search, retrieve, and display (Harold, 2005). Selecting storage solution for TBX-format data is complicated due to the hierarchical data structure and many optional data categories. Storage solution directly affects efficiency of data search, retrieval and display.

There are several of options for storing XML compliant TBX data. Three options analyzed are: storing data in relational database, storing XML data in a file system, and storing XML data in a relational database. Each of these solutions was analyzed in accordance with previously mentioned data management tasks.

Keeping data only in XML format would increase complexity of data retrieval operations and would negatively affect performance of search. At the same time storing XML data in a relational model makes it technically difficult; this model is not flexible enough for varying structure of terminological data. The data model must provide a possibility to store all possible terminological data categories, but it also must take into account that most of them will usually not be filled.

Storing hierarchical data in a relational model leads to the extraction of data categories before it is possible to store XML data in a database, because every data category must be stored in a separate field. Database fields have a limited size while the XML data structure field may have a virtually unlimited size. While storing data intact in XML format solves these problems, new problems arise, for example, if data is stored in XML – then data may be validated only with XML schema. When storing XML data in database intact, there is a possibility to duplicate only the required data categories. This leads to some form of information extraction, but this extraction is limited to a number of data categories that have a lower probability of changing over time.

After data has been stored, problems of data retrieval must be analyzed. As the TBX format is devised for terminological data exchange, it is used not only as the unified format to which terminology is transformed, but also as an exchange mechanism with other systems. So the data from the data source must be retrieved in the very same TBX format. If data is stored in the relational model, data categories from numerous fields must be merged into a single TBX compliant XML data structure. In this case the process of creating terminological entry equivalent to the original takes lot of processing power and affects system performance. However, when storing data in XML format, extraction is rather simple and accurate as no data transformations have to be made.

Search is used in terminological systems to retrieve entries. To create a user-friendly system, search response time must be reasonable and acceptable to users. This is one of the most important things to consider when choosing the storage solution. Standard guidelines for ideal web response times are (Nielsen 1999):

- 0.1 second. Ideal response time. The user does not sense any interruption.
- 1 second. Highest acceptable response time. Download times above 1 second interrupt the user experience.
- 10 seconds. Unacceptable response time. The user experience is interrupted and the user is likely to leave the site or system.

Search results were analyzed based on prototypes for every storage solution. It was discovered that the average response time for search in XML files stored in a file system is 5 seconds. Search in the relation model took about 0.01 second, which was the same result as for the model where XML data is stored intact in a relational database and the data categories for search and retrieval are duplicated in separate fields (Liedskalniņš, 2007).

For storing large volumes of multilingual terminology data we recommend the realization model where XML data is stored in a relational database and data categories necessary for data management tasks are extracted and stored in separate fields. Applicability of this model has been proven by implementation in EuroTermBank termbase. This model shows optimal results for all data management tasks. While it requires additional processing and data preparation when storing data, it provides excellent data retrieval and search.

4. TERMINOLOGY DATA MODELING PRINCIPLES

According to (ISO 26162, 2010) data modeling is a process of structuring and organizing data, typically for implementation in a database management system. This chapter describes our recommendations for modeling of terminological data for terminology consolidation projects.

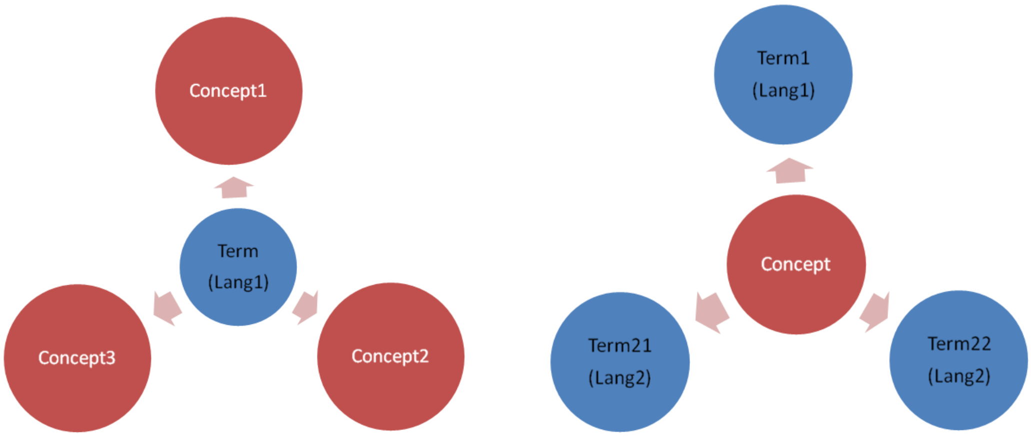


Figure 9 Transition from lexicographical approach to concept-oriented approach

4.1 APPLICATION OF STANDARDS IN SCENARIO BASED DATA MODELING

To ensure exchangeability and facilitate recognition and comprehension of data categories for new or outside users terminology data should be modeled based on ISO 12200:1999 and 12620:1999 standards. The principles of these ISO standards require that the term entries are concept oriented, contain a rather broad selection of data categories that permits the necessary level of detail and permit full descriptions of each term.

However, these standards are very extensive and general and there is a strong need for guidelines on how to apply them in particular usage context or applications.

We propose to apply scenario based approach in terminology data modeling. It is based on three scenarios we introduced in Chapter 3.2 - local, national and international terminology work scenario. In this chapter we will describe application of this data modeling approach basing on (Henriksen et al., 2006).

4.1.1 DATA MODELING IN LOCAL SCENARIO

Within the local scenario, the main conditions and goals that are important for the design of a data structure are: tight time frames, translation-oriented needs, exchangeability, and limitation of terminology work to one or a few domains. These criteria speak in favor of a highly customized and only moderately exhaustive data structure where data categories are consistent with the requirements of the particular application area and have a translation related focus.

A focus on translation requirements implies coverage of more than one language. It must, therefore, be considered whether such descriptive concept related information as definitions or explanations are necessary for each language or only for one language. If the term collection is multilingual, a definition for each language is usually necessary. If the term collection is only bilingual, it may not be necessary.

A focus on translation requirements also indicates inclusion of data categories permitting sufficient information about the use of a term, for example, different types of grammar information, context information and collocation information. Some translation settings may also require grammar information for each word of a term. Furthermore, it is often considered very important to document the degree of equivalence between terms of different languages. Data categories that could be relevant in this respect are, for example, false friend, directionality and transfer comment.

The below data structure containing four levels reflects a multilingual terminology setting permitting, for example, concept descriptive information for each language and grammar information for each word. In multilingual as well as bilingual terminology settings it can however be considered to omit the word level and locate grammar information at the term level instead. In some bilingual terminology settings it can also be considered to have a definition for only one language. Consequently, the data structure in a bilingual framework may include only 2 levels, namely, concept and term levels.

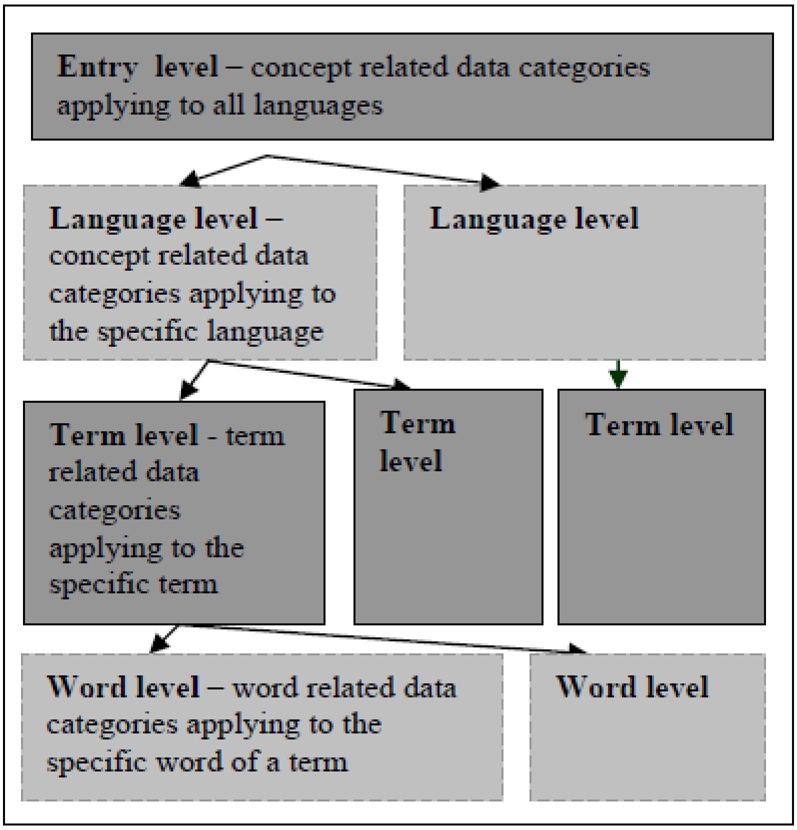

Entry level – concept related data categories applying to all languages
Language level – concept related data categories applying to the specific language
Language level
Term level - term related data categories applying to the specific term
Term level
Term level
Word level – word related data categories applying to the specific word of a term
Word level

Figure 10 Four-level structure for terminology data

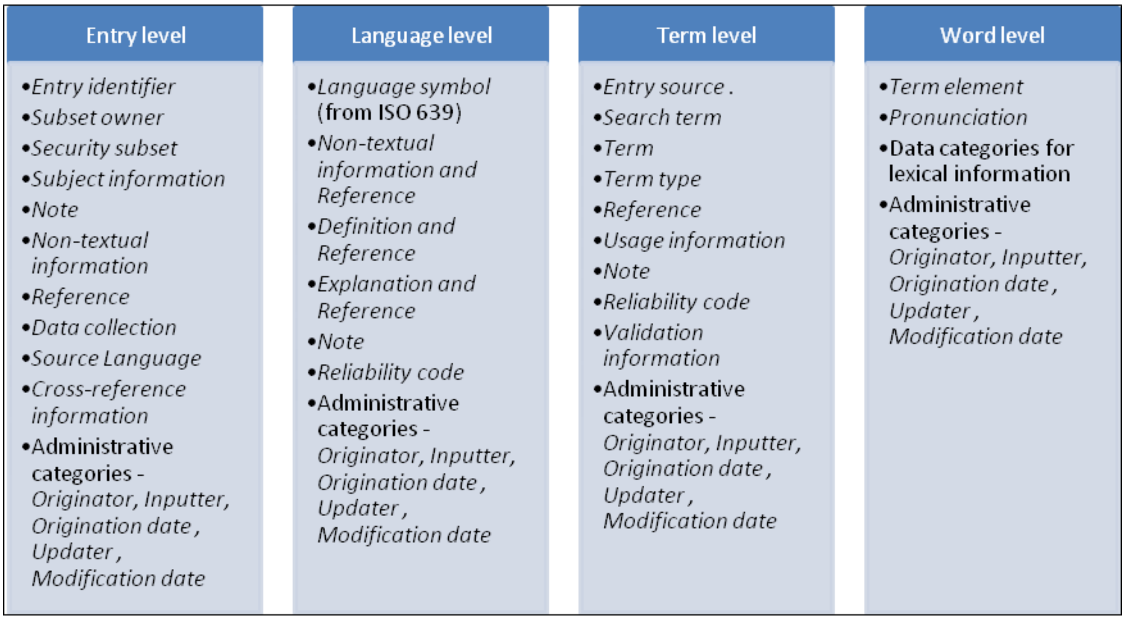
4.1.2 DATA MODELING IN NATIONAL SCENARIO

In the national scenario, conditions and goals influencing the design of a data structure are adequate financial support, exchangeability, broad domain coverage and high quality in general terms. Besides, a national term collection is aimed at terminology coordination and regulation rather than at translation. These criteria point towards a data structure that permits an exhaustive selection of data categories covering very different user requirements and enabling users to develop entries for very different purposes and of a very high quality.

This implies that the data structure should often contain 2 levels: concept and term levels (at least when the term collection is monolingual) and that data categories should represent a wide selection of information types and include term status qualifiers reflecting for example acceptability, approval or applicability of a term in a given context. An example of a term status qualifier is normative authorization which is assigned by an authoritative body and includes qualifiers as standardized term, preferred term, admitted term and deprecated term.

4.1.3 DATA MODELING IN INTERNATIONAL SCENARIO

Within the international scenario, the criteria considered important are very similar to those important in a national scenario. A crucial difference is however that international terminology cooperation is multilingual by nature. Therefore it is recommended that the data structure should include four levels permitting concept descriptive information for each language and grammar information for each word of a term.



Entry level

- *Entry identifier*
- *Subset owner*
- *Security subset*
- *Subject information*
- *Note*
- *Non-textual information*
- *Reference*
- *Data collection*
- *Source Language*
- *Cross-reference information*
- Administrative categories - *Originator, Inputter, Origination date, Updater, Modification date*

Language level

- *Language symbol* (from ISO 639)
- *Non-textual information and Reference*
- *Definition and Reference*
- *Explanation and Reference*
- *Note*
- *Reliability code*
- Administrative categories - *Originator, Inputter, Origination date, Updater, Modification date*

Term level

- *Entry source.*
- *Search term*
- *Term*
- *Term type*
- *Reference*
- *Usage information*
- *Note*
- *Reliability code*
- *Validation information*
- Administrative categories - *Originator, Inputter, Origination date, Updater, Modification date*

Word level

- *Term element*
- *Pronunciation*
- Data categories for lexical information
- Administrative categories - *Originator, Inputter, Origination date, Updater, Modification date*

Figure 11 Proposed data model for international scenario based on ISO 12620 data categories.

4.2 EUROTERMBANK: CASE STUDY FOR DATA MODELING IN INTERNATIONAL SCENARIO

Analysis of user needs and existing standards provided the basis for development of data structure for EuroTermBank system and recommendations for developers of different terminology resources. Irrespective of type of organization, purpose of terminology and type of domain, it is as a principal rule recommended that the data structure permits a broad selection of data categories that provides exhaustive list of information types and enables users to develop entries of a high quality.

It will usually be recommendable to develop a data structure of 2 to 4 levels dependent on the number of languages involved. If the term collection is monolingual, it is recommended that the data structure contains 2 levels; one level for conceptual information and one level for term related information. Examples of conceptual information are *domain*, *definition* and *explanation* and examples of term related information are *term* and *context*. This data structure will allow many terms to designate one concept (one definition).

For a bilingual term collection 3 levels may be used that apart from conceptual and term related information also permit lexical information of the individual words that constitute a term. Whether a word level should be added to the data structure depends mostly on the nature of the foreign language.

A consequence of the data structures described above is however that a definition in only one language can be created for each concept. This may constitute a problem in a bilingual term collection as users may not speak the same native language (and therefore may not fully understand the definition) and as minor conceptual differences must sometimes be expressed. If definitions in both languages are requested, it is recommended to split the conceptual level in two: One level for the language specific information (*language level*) permitting, for example, a definition for each language and one level for the language independent information (*entry level*) containing, for example, domain information.

The Table 3 shows data categories present in all the data sets.

Data category	*Information type*	*Description*	*Comments*
Term related	Term	Language 1	
Language related	Definition	Language 1	
Language related	Context/example	Language 1	
Language independent	Subject field	Implicit/explicit	When a term collection covers only one domain, this info is implied
Administrative	Administrative	Serial number/ entry identifier/ author code, etc.	Various subtypes are used

Table 3 Data categories present in all EuroTermBank term collections.

It is essential that the data structure is based on standards to ensure exchangeability with other data collections and to ensure that data categories are recognizable for outside users. Terminology data structure should comply with ISO standards 12200 and 12620 (this is recommended for all levels). The original ISO 12620 was designed specifically for concept oriented terminology management systems but it is also targeted for a broader usage in different terminology applications.

EuroTermBank data structure comprises information about the concept, the terms that designate the concept, and the words that constitute the individual terms. As a multilingual system it should permit definitions in all languages and therefore conceptual information should be grouped in two levels: the *entry level* containing language independent information and the *language level* containing language specific information. Term related information should be contained at *term level*; an example of an information type that might appear at term level is usage information. Lexical information concerning a specific word should be contained at *word level*.

The data structure developed for EuroTermBank comprises up to 4 hierarchical levels based on ISO standards 12200 and 12620, as described in detail by (Wright, 2005):

The **entry level** provides concept-related data categories applying to all languages. It contains language-independent information like entry identifier, subject information, data collection; administrative information like subset owner identifying the institution responsible for the entry; originator, origination date, updater, modification date and a number of other fields.

- *Entry identifier* – The value of this data category is a system-generated number that will identify the entry uniquely.
- *Subset owner* – The value of this data category is the institution responsible for the whole entry. As the data collection will contain contributions from many different organizations it is necessary to state clearly who is responsible of maintenance of each entry.
- *Originator* – An identifier of the person who prepared the entry.
- *Inputter* – An identifier of the person who types in the information.
- *Origination date* – The date the entry was first created.
- *Updater* – The value of this field is the person having made the latest changes to the information at entry level.
- *Modification date* – The date when the latest changes to the entry level were made.
- *Security subset* – This data category contains a security classification expressing the confidentiality level of the entire entry. A security classification can be used in connection with for example critical terms during a development phase
- *Subject information* – The data category(ies) chosen for subject information will contain the domain of the particular concept.
- *Note* – A free descriptor field to allow for other kinds of subject information that cannot be expressed in the subject information field(s).
- *Non-textual information* – The data category(ies) chosen for non-textual information will contain for example tables, figures, videos and other binary data.
- *Reference* – Reference(s) to the non-textual information.

- *Data collection* – This field can be used to signify that a particular concept belongs in a particular collection of concepts.
- *Source Language* – This information concerns the source language of a set of terms that are not perfectly multi-directional. There is currently no 12620 data category to indicate the source language in a set of terms that are not perfectly multi-directional, but there are some alternative possibilities that can be considered.
- *Cross-reference information* – A reference to other concepts in various ways related semantically to the concept in question, for example broader concept, subordinate concept or related concept.

Language level. Provides concept-related data categories applying to the specific language. It contains language-specific information like definition, reference, explanation and others, as well as administrative information.

This level can contain the following administrative fields – *Originator*, *Inputter*, *Origination date*, *Updater*, *Modification date* (see descriptions above).

Language symbol – This data category contains the language symbol of the particular language. The symbols specified in ISO 639 should be used.

Non-textual information and Reference – See comment about non-textual information at entry level.

Definition – In this field, a formal and precise description of the concept is given.

Reference – Reference(s) to where the definition given above was found.

Explanation – Compared to the *Definition* field, this field makes it possible to give a more informal description of the concept. This field would be particularly useful in cases where a formal definition has not been obtainable.

Reference – Reference(s) to where the explanation given above was found.

Note – This data category can contain some additional and general information about the concept in the particular language or the field can contain information related to the definition or explanation.

Reliability code – Reliability codes are suggested at language and term levels. A reliability code at the language level will thus provide an assessment of the correctness and precision of the information given in relation to the specific concept.

Term level. Provides term-related data categories applying to the specific term. It includes term-related information like term in a particular language, entry source, search term containing related forms of the term to facilitate search, reference with source(s) of the term, usage information, and others.
Originator, *Imputer*, *Origination date*, *Updater*, *Modification date* – Contains the same information as in levels above.

Entry source – If the entry is imported from another resource this field will always contain information about the database or format from which data are imported.

Search term – This field will contain related forms of the term to facilitate searching. The author of term level information containing a verb may e.g. expect that users will often make a search for the adjectival form. In this case the author can state the adjectival form in *search term*.

Term – This field will contain the term: a designation of a defined concept in a specific language by a linguistic expression.

Term type – The value in the *Term Type* field is an attribute assigned to a term. The values can be selected from a picklist containing the term types used by the organizations. A picklist for *termtype* is contained in ISO 12620.

Reference – Source(s) of the *term.*

Usage information – Data categories selected for usage information may for example concern a textual example of a concrete use of the term in question, a classification indicating the relative level of language of a term, information about the use of a particular term over time, the status of a term with respect to standardization etc.

Note – A general comment that applies to the entire term level.

Reliability code – Reliability codes are suggested at language and term levels. A reliability code at the term level will thus provide an assessment of the correctness and precision of the information given in relation to the specific term.

Validation information – It is suggested that validation information is located at term level and not at the other levels though a validation procedure includes validation of all levels.

Word level. Provides word-related data categories applying to the specific words of a term. A term may be a multiword string, therefore, this level is created to contain lexical information that concerns the individual words of a term. Data categories for lexical information are, for example, part of speech, grammatical number, grammatical gender etc.

Below the data categories are described that are suggested for each level of an entry. The organization of data categories is by level, i.e. if a data category can appear at several levels, it is repeated for each of these levels. Although these data categories comply with ISO 12620 this is by no means an exhaustive list of ISO 12620 data categories as standard contains multiple possibilities that must be considered in relation to the specific application.

As a term may be a multiword string this level is created to contain information that concerns the individual words of a term.

This level can contain the following administrative fields – *Originator*, *Inputter*, *Origination date*, *Updater*, *Modification date*.

Term element – This data category concerns a particular word that forms part of a term.

Dependent on the languages involved in the international cooperation some data categories for grammar information should be selected. Data categories for lexical information are for example, part of speech, grammatical number, grammatical gender etc.

Depending on involved languages and purpose of terminology, pronunciation information may be necessary.

5. FEDERATION PRINCIPLE IN TERMINOLOGY CONSOLIDATION

The rapid development of the internet created an interest to apply its capabilities to the problem of fragmentation of terminology resources on the internet across diverse term banks and terminology projects. A number of user scenarios require consolidation on a multilingual and multinational scale.

The necessity to move away from the single, isolated data bank towards a multi-bank environment was suggested by (Cabré, 1999). She suggested to simultaneously access several data banks that are all integrated into an overall working structure that not only includes the databases, but also other computerized tools and resources.

The federation of term banks is a new concept in linking portals and data repositories, and it should go far beyond the establishment of pointers or links, towards the level of exchangeability and semantic interoperability of data and data structures (Galinski, 2007).

The author has proposed a federated approach to consolidate distributed terminology resources in (Vasiljevs & Rirdance, 2007). It should be noted that while researching previous work in this area, the author found the first mention of terminology federation in presentation slides of (Hodge, 2000) where he proposes federation as a possible solution for thesaurus interoperability in digital libraries.

The federation principle provides unified access and consolidated representation of distributed multilingual terminology data from various institutions in different countries. Individual systems have their own system architecture, data structure and user interface, but they are dynamically interlinked using standards-based terminology exchange mechanisms.

To ensure the viability of the federated system of terminology databases, inclusion of a term bank in this federated model requires it to be independently supported and maintained both institutionally and technically.

The federation model can be applied on all terminology work levels but the main benefit from this model is for international terminology work for international organizations and global companies and projects.

The key approach for consolidation is the federation principle where multiple autonomous databases are transparently integrated into a virtual federated database. The federated approach to terminology consolidation provides a solution to at least one inherent challenge of all terminology banks – maintenance of terminology is done at the local or national level, and the changes at the local or national level become instantaneously available for integration with other federated resources.

EuroTermBank implementation of federation demonstrates that terminology consolidation through database federation at an international level unites dispersed terminology databases from different countries into single terminology portal.

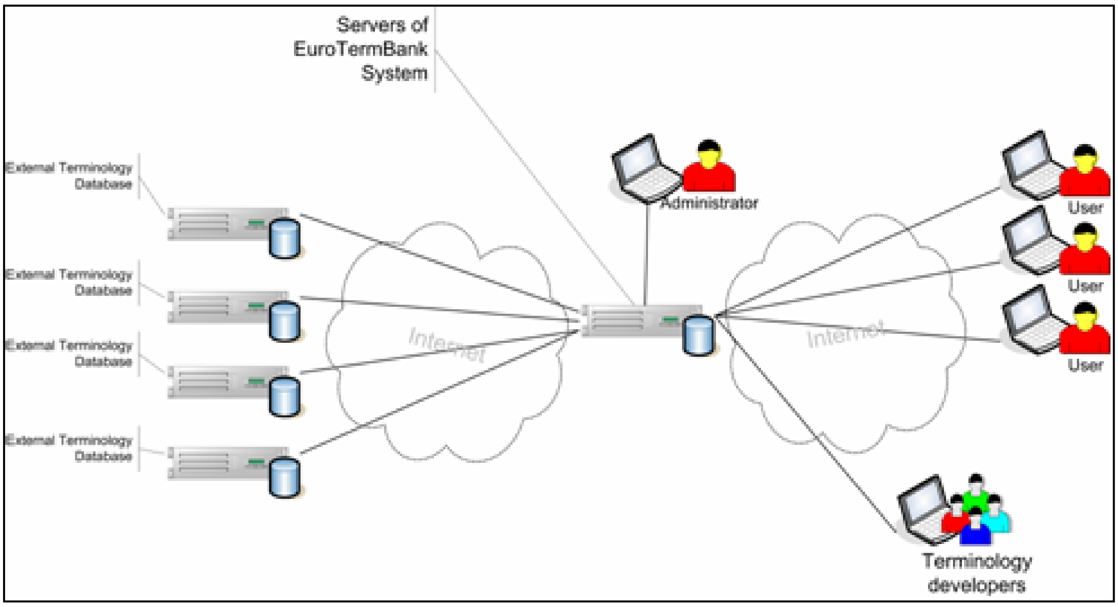

Servers of EuroTermBank System
External Terminology Database
External Terminology Database
External Terminology Database
External Terminology Database
Internet
Administrator
Internet
User
User
User
Terminology developers

Figure 12. Schematic model of federation principle as applied in EuroTermBank.

An important by-product of this approach is the promotion of a unified methodology for terminology work and application of industry standards.

The federation of term banks has the potential of becoming a key concept in presenting terminology resources in a user-friendly way, as it provides a single meta-search interface to a number of interconnected, or federated, term banks.

Currently, EuroTermBank provides federated access to several interlinked term banks:

- IATE – terminology database of European Union institutions;
- Termnet.lv - Official Latvian terminology database of Terminology Commission of LAS;
- PolTerm – Polish legal terminology database;
- Hungarian legislation terminology database.

The major federated term bank is IATE, the inter institutional terminology database of the EU (Rummel, Ball, 2001). As IATE is the most used online terminology database (Gornostay, 2010) it was a natural choice to interlink it with EuroTermBank providing single access point for terminology research. After interlinking with

EuroTermBank, query results from IATE database are available from the EuroTermBank portal.

The federated system provides the user with a single access point to a vast array of terminology data. The model of online presentation of federated resources from multiple term banks is relatively new and puts forward a number of challenges, from the expected level of quality of terminology resources and maintenance patterns of each interlinked term bank to the user interface issue of presenting multiple partially overlapping entries across a number of federated sources.

The federation model also poses issues related to ensuring reliability of the sources or of the source data in case an important resource of the federation becomes unavailable, temporarily or ultimately, and ensuring a unified approach to change management on all levels, from data structure to the changing terminology content and preservation of legacy data. Another common challenge to terminology termbanks exacerbated in a federated model is the application and mapping of subject field classification systems. A major challenge is the implementation of a concept-oriented approach requiring a certain level of concept harmonization in a multilingual setting with diverse terminology creators.

However, these challenges are inherent to all terminology work, even on individual level.

EuroTermBank’s advantage lies in a more efficient and consolidated approach in solving these challenges, compared to the uncoordinated and oftentimes partial and incompatible solutions typical at the local level.

5.1 IMPLEMENTATION OF THE TBX STANDARD

TBX (TermBase eXchange) is an open XML-based standard format for terminological data, created to facilitate interchange among termbases. This standard provides a number of benefits as TBX files can be imported into and exported from most software packages that include a terminological database . For interoperability of terminological data, it is important to use open standards for data exchange. TBX as XML-based standard also provides platform-independent data exchange. It is intended to qualify as a TML (Terminology Markup Language), as defined in the TMF (Terminology Markup Framework) specified in ISO 16642:2003. In addition, TBX is intended to support the extraction and merging of information from other, non-TMF-compliant, formats, although these processes may involve some information loss. Besides TBX tags, the TBX format may include also meta-information tags, which allows including such information as HTML formatted data.

TBX standard is based on three ISO standards: ISO 12620, ISO 12200 and ISO 16642. ISO 12620 defines data categories to be used for terminological data storage either in digital or printed format. Terminological data categories described in this standard are divided into three large subgroups that contain more detailed data category sections:

- Term-related information
- Descriptive information
- Administrative information

As a standardized exchange format, TBX can be used as the interchange format between single system components. Moreover, it facilitates terminological information interchange among termbases with different data models, thus improving interoperability of terminological data globally

According to the hierarchy of a TBX document, the highest-level XML element is the *martif* element, which contains a <*martifHeader*> element and a <*text*> element. The <*martifHeader*> element provides a description of the file, on the applicable XCS file and unusual character encoding, and a history of major revisions to the collection.

The <*text*> element contains the terminological data. It includes in the <*body*> the actual terminological entries – one entry per concept – enclosed in <*termEntry*> tags, as well as complementary information, e. g. bibliographical data, in the <*front*> and <*back*> elements, to which can be referred from the <*body*> entries. Within the terminological concept entries various data categories allow to provide different kinds of information, either in free text or chosen form a pick list, as well as cross-references that points to either somewhere inside the *martif* element or to an external object using a URL. The terminological concept entries (<*termEntry*>) can be multi- or monolingual.

Concepts in terminology correspond to objects in the real world. Concepts are mental constructs functioning as 'first order representation', whereas the corresponding terms (or other kinds of concept representation) have the role of ‘second order representation’ (Galinski, 2005).

```
<?xml version='1.0'?>
<!DOCTYPE martif SYSTEM  "./TBXcoreStructureDTD-v-1-0.DTD">
<martif type='TBX' xml:lang='en' >
<martifHeader>
      <fileDesc>
            <sourceDesc><p>from an Oracle corporation termBase</p>
            </sourceDesc>
      </fileDesc>
      <encodingDesc><p type='DCSName'>TBXdefaultXCS-v-1-0.XML</p>
      </encodingDesc>
</martifHeader>
<text> <body>
      <termEntry id='eid-Oracle-67'>
            <descrip type='subjectField'>manufacturing</descrip>
            <descrip type='definition'>A value between 0 and 1 used …</descrip>
            <langSet xml:lang='en'>
                  <tig>
                        <term tid='tid-Oracle-67-en1'>alpha smoothing factor</term>
                        <termNote type='termType'>fullForm</termNote>
                  </tig>
            </langSet>
            <langSet xml:lang='hu'>
                  <tig>
                        <term tid='tid-Oracle-67-hu1'>
                              Alfa sim&#x00ED;t&#x00E1;si t&#x00E9;nyez&#x00F5;
                        </term>
                  </tig>
            </langSet>
      </termEntry>
</body> </text>
</martif>
```

Table 4 Example of a TBX document.

TBX includes meta-markup tags for distinguishing embedded non-TBX markup from text. They allow TBX elements to contain various kinds of other markup, e. g. html or text processing markup that needs to be retained but should not necessarily be processed during terminology management functions.

EuroTermBank system implements the TBX standard to satisfy a number of requirements: enabling data exchange between different ETB modules, interoperability with external databases, data import/export, and data storage in the EuroTermBank internal database.

A list of required terminological data categories was created during the EuroTermBank project based on best practice research. Selected data categories were compared to data categories specified in ISO 12620 to verify their compatibility. As TBX standard defines XML-based format, it was possible to use only the required data categories and still be compatible with TBX standard.

Although TBX standard is mainly devised as an exchange format, in EuroTermBank it is also used for terminological data storage in the database, as terminology has specific characteristics that make it difficult to store such type of data:

- It has many optional data categories;
- Data categories frequently have no format restrictions;
- Size of some data categories is not predictable.

These problems were solved in EuroTermBank by storing data in the XML-based format defined in the TBX standard. This provided the following benefits:

- Storage of all TBX data categories;
- No format and size limitations for data categories;
- Simple extensibility.

The TBX standard is also used for data import and export to and from EuroTermBank database. All resources to be included in the portal’s internal database are converted to TBX format. Source formats vary from resources to highly structured XML files. As TBX is also the storage format, there are no significant reasons for introducing another format. As TBX allows storage of all standardized data categories, it is possible to convert all resources to TBX format. Even if resources have resource specific data categories that are not included in the standard, it is possible to store these categories as supplemented XML tags without changing the physical data storage model.

TBX format is applied throughout the EuroTermBank system. Since TBX format is used through all the resource life-cycle stages, it also ensures data consistency. Using an open and non-proprietary standard is appropriate not only for EuroTermBank resource interoperability within the internal system, but also for communicating globally with external terminology databases. EuroTermBank system is designed to provide external systems with standardized data in TBX format and receive data from external systems in the very same way. There is no need to define a new framework either for processing every single external data provider or for the data provided by the system.

In the EuroTermBank system, the TBX standard enables data storage of all four terminological concept levels – entry level, language level, term level and word level (Schmitz, Vasiljevs, 2006). It also supports all data categories identified during the best practice research. All of 92 resources imported into EuroTermBank have been converted to TBX format without data loss, ensuring not only standard compliance, but also extensibility of the format.

Using the TBX standard throughout the system provides data consistency as data are not converted either in the system's internal modules or in the communications with external systems. From external systems that are already connected to the EuroTermBank system, one is directly providing data in TBX format. Other systems use proprietary exchange formats so conversion to TBX is applied before passing data to EuroTermBank. Furthermore, there are several systems that are on the way to use EuroTermBank system as the data source for terminology and communicate in the TBX standardized format.

5.2 LIMITATIONS OF TBX

Taken its strength in terminological data storage and exchange, TBX also has some weaknesses in data interoperability. TBX does not solve the problem of interoperability that originates from the application of different data categories across term banks, for example, some data categories might be required in one termbase, while optional or not present in another one, or one and the same data category may appear on different levels of the entry structure. Also, there is no straightforward way for creating relations between terminological entries from different resources. Although technically it is possible, it is not part of the standard. The situation in creating relations between single resource entries is a bit better; a few types of relations – broader, generic and related – are defined there. However, these relations are limited and would be insufficient for creating more complex ontology structures.

TBX falls short of ensuring blind interchange between any given implementations, since it provides ample freedom, for example, in application of data categories. Thus some data categories may be required in one term bank, while optional or not present in another one, or one and the same data category may appear on different levels of the entry structure. Although TBX is not intended to ensure blind interchange, this limitation hampers its wider implementation.

Therefore an important step forward is development of TBX-Basic, a lightweight version of TBX that identifies a limited set of data categories, including a minimum set of mandatory categories (www.lisa.org/sigs/terminology). It is meant to satisfy the requirements for small or medium sized language industries and will be included in TBX as an appendix demonstrating an example of a TML (Terminology Markup Language) that is compatible with TBX.

TBX is also criticized for its concept-based multi-linguality and non-directionality, stating that TBX does not cover terminology in areas that are subject to societal or cultural influences and where there is no concept with synonymous terms in many languages (Thurmair, 2006). Thurmair concludes that TBX is only suitable for the representation of technical terms where a 1:1 correspondence between participating languages can be assumed.

In response to this criticism we should take into account that TBX does provide for language-specific descriptions of concepts using definition, comment, context or other text field. In cases where 1:1 correspondence is not present, a new concept with either only one or a limited set of languages can be defined. While it is true that TBX is not suited for exchange of machine translation dictionaries that contain a large number of general vocabulary terms, this is not the purpose of TBX. As shown by EuroTermBank experience, TBX does serve as a practical and highly usable exchange format for a number of terminology exchange scenarios.

The concept of terminology exchange becomes relevant and important in scenarios involving merging or exchange between several terminology resources or collections, which involves collating or merging term entries across collections as described in this article. Despite this being a major scenario in terminology exchange, there is no straightforward way in TBX for creating relations between terminological entries from different resources. Although technically it is possible, it is not part of the standard. The situation in creating relations between single resource entries is better, with a few types of relations – broader, generic and related – explicitly defined within the standard. However, these relations are limited and would be insufficient for creating more complex ontology structures.

6. TERMINOLOGY DATA REPRESENTATION AND SHARING

A number of key terminology resources on the web have undergone substantial changes in recent years, and new ways of reaching their users have emerged. We will identify and describe the new patterns of presenting terminology resources online, taking EuroTermBank as the basis.

The general paradigm shift in the usage of the World Wide Web that is somewhat vaguely referred to as Web 2.0 (Musser & O'Reilly, 2006) has impacted the area of terminology resources as well. Terminology users today expect much more than a static database with a few search options. In representation of terminological data we focus on such notions as user-centered design, consolidated representation through entry compounding, data sharing patterns, interoperability and user participation in term banks.

The author argues that, for a successful operation of a term bank, today's imperative is reaching out for the user and delivering the required content, wherever it may reside, with the method and in the format required by the user. The area of user participation and interaction is identified as yet to be successfully integrated in the design of terminology portals.

6.1 TERMINOLOGY ENTRY COMPOUNDING

Entry compounding solves the problem of unified representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. The majority of terminology resources that are available in Eastern European countries are bilingual with a source language mostly being English. Much smaller number of resources are monolingual or have terms in three or more languages.

Since multiple terms in multiple languages can refer to the same concept, the concept is the shared element that must be used to link the terms together in a multidimensional database (Wright, 2005).

In previous chapters we strongly advocate to model data structure according to a concept-oriented approach to terminology. Terminology entry denotes an abstract concept that has designations or terms as well as definitions in one or more languages.

If a terminology bank contains entries coming from different collections and designating the same concept we have an obvious interest to merge them into one unified multilingual entry.

For example, if we have a term pair *EN computer – LV dators* coming from Latvian IT terminology resource and another term pair *EN computer – LT kompiuteris* from Lithuanian IT terminology resource we may want to join these two into unified entry *EN computer – LV dators – LT kompiuteris.* Such multilingual entry allows to get correspondence between language terms that are not directly available in any terminology resource (in our example new term pair *LV dators – LT kompiuteris).*

But merging entries just on the bases of a matching term in one language that is common for these entries will lead to many erroneous term correspondences. For example, if we have LV-EN entry *stumbrs-stick* and ET-EN entry *kang-stick*, we may want to merge these entries into compound entry LV-ET-EN *stumbrs-kang-stick*. But if we would add to this alignment LV-EN entry *rokturis-stick* it would lead to wrong LV-ET translation *rokturis-kang*.

Such problems are obviously due to the frequent ambiguity of terms among subject fields or rarer cases of ambiguity in the context within one subject field. We can conclude that the only error-free method for merging entries is evaluating whether these entries denote the same concept. Unfortunately in practice it is often impossible or very expensive to make comparisons of cross-lingual terminology concepts. There is a lack of experts with sufficient knowledge of respective languages and subject fields. The task is considerably hindered by the fact that majority of EuroTermBank terminology collections do not have term definitions included.

To solve these problems we propose a new method for consolidated data representation – the *terminology entry compounding*. Entry compounding is an automated approach for matching terminology entries based on available data.

The most reliable indication for matching entries is having unique and unambiguous concept identifiers. The best example is terms from ISO terminology standards. These term entries have an identifier in the form *[Standard_identifier].[term_number]*. Accordingly, all national standards share the same identifier for corresponding entries and can be merged with a very high degree of reliability. Another case of unique internationally applied identification is the usage of Latin names in medicine and biology (with a number of exceptions with different Latin names designating the same concept). If there is no unique identification for concepts in collections, less precise matching criteria are used, namely, the English term and the subject field. English was chosen as the most popular language in term resources.

EuroTermBank uses Eurovoc as a subject field classification. A number of terminology resources use only top classification levels of Eurovoc but there are many resources with detailed classification using Eurovoc sublevels of different depth. For this reason it was decided to take into account only the top classification level for entry compounding. This means that sublevels are equalized to the top classification level.

It is important to understand that entry compounding is a data representation method that does not propose to create new terminology entries. It is a visualization aid that displays matching entries across collections in a consolidated way. Matches are determined by applying a number of criteria and as such cannot be error-free.

As majority of terminology resources integrated in EuroTermBank are bilingual (Table1), we would like to transform data representation from number of separate bilingual entries to unified multilingual record.

Entry languages	*Number of entries*	*Percentage from total*
monolingual	11230	2%
bilingual	398854	68%
3-lingual	45497	8%
4-lingual	69134	12%
5-lingual	48761	8%
>5-lingual	12216	2%

Table 5 Multilinguality of EuroTermBank source records

Entry compounding solves the problem of visual representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. At present, the EuroTermBank database contains over 585,711 term entries with more than 1,500,500 terms. When applying entry compounding, over 135,000 or 23% of entries get compounded. Hence entry compounding is a considerable aid for the user in finding the required term, for example, in the translation scenario between language pairs for which term equivalence is not established in existing collections.

Unfortunately abovementioned criteria for entry compounding are insufficient and generate too much incorrect alignments. A high recall rate also leads to relatively low precision although we currently do not have exact precision evaluation figures.

If our term entries would include term definitions then we could compare these by human review or by applying automated analysis methods. But because large majority of Eastern European terminology resources do not include definitions we need to look for other sources to depict meaning of terms.

We suggest using multilingual text corpus as a source were to look for term usage patterns and attempt to disambiguate its meaning. Of course it is impossible to get term definition from the regular text corpus. But we can intuitively assume that term meaning is related to the context where term usually appears in. This intuition has also some rational basis. For cost and time saving many institutions dealing with terminology creation are not preparing definitions for new terms but instead include in term database several typical examples of usage context.

We can assume that term t in language L1 and s in language L2 are matching (or denoting the same concept) if t and s have similar context patterns in L1 corpus and L2 corpus respectively. By the context pattern we mean characteristic collocates frequently appearing in proximity of term. Because terminology is related to special language (special language uses specific words with specific, preferably unambiguous meaning, in contrast to general language with wide lexicon of usually very ambiguous words) we are interested in those collocate words that are terms from the same subject field. This is also based on common intuition that term in specific subject field should be best described by other terms from this subject field.

6.1.1 PROPOSED METHOD

In the proposed method we try to grasp the intuition that if two terms in different corpora have similar context patterns then they might denote the same concept and more frequent collocations have more impact on term context pattern than less frequent ones.

Let's assume that we have applied simple term compounding for bilingual terminology resources as described previously. For language *L1* term *t* we have several translation candidates s_1, s_2, ..., s_n in language *L2*. Our task is to select the most probable from these candidates by analyzing context patterns of these terms.

Let's denote frequency of term *t* in language *L1* corpus with *count(t)*.

Frequency of s_1, s_2, ..., s_n in *L2* corpus will be denoted with $count(s_1)$, $count(s_2)$, ..., $count(s_n)$.

We denote collocations of term *t* with $coll_1(t)$, $coll_2(t)$, ..., $coll_m(t)$ and respective frequency of these collocations in proximity with *t* with $count(t, coll_1(t))$, $count(t, coll_2(t))$, ..., $count(t, coll_m(t))$.

We will select those collocations of the term *t* in language *L1* whose frequency is higher than certain threshold *p*.

This means that we will select $coll_j(t)$, where $\frac{count(t, coll_j(t))}{count(t)} > p$.

For every such collocation we will find translation candidate $x_1, x_2, ..., x_k$ in language *L2*. For every candidate translation s_i of the term *t*:

if $\frac{count(s_i, x_1) + count(s_i, x_2) + ... + count(s_i, x_k)}{count(s_i)} > p$ then we will add to the score of this candidate the lowest from the numbers

$$\frac{count(s_i, x_1) + count(s_i, x_2) + ... + count(s_i, x_k)}{count(s_i)} \text{ and } \frac{count(t, coll_j(t))}{count(t)}.$$

Now let's do the same calculation from reverse side – for every translation candidate si in language L2 we will select collocations whose frequency is higher than certain threshold p.

This means that we will select $coll_j(s_i)$, where $\frac{count(s_i, coll_j(s_i))}{count(s_i)} > p$.

For every such collocation we will find translation candidates $x_1, x_2, ..., x_k$ in language *L1*.

If these translations appear in context with t frequently enough passing our threshold p:

$\frac{count(t,x_1)+count(t,x_2)+...+count(t,x_k)}{count(t)}$ > *p*, then we will add to the score of this candidate the lowest from the numbers

$\frac{count(t,x_1)+count(t,x_2)+...+count(t,x_k)}{count(t)}$ and $\frac{count(s_i,coll_j(s_i))}{count(s_i)}$.

We will assume that translation candidate s_i with the highest resulting score is the most probable equivalent of term *t* in language *L2*.

6.1.2 EXPERIMENTAL RESULTS

To test the proposed method we carried out an experiment on compounding Latvian and Lithuanian terms. For this experiment we used JRC-Acquis Multilingual corpus v3.0 which is the largest publicly available source of corpus data for Latvian and Lithuanian (Steinberger et al., 2006). Latvian corpus contains 22 906 documents with 27 592 514 words. Lithuanian corpus contains 23 379 documents with 26 937 773 words.

It could be asked why not to use well proven statistical alignment methods to align terms from these corpora as these are highly parallel texts mostly being translations from the same source (English). But as we want to find a method for more general case of lack of parallel in-domain data, we split this corpus in 2 parts. For the Latvian corpus we used the first part and for Lithuanian – the second. In such a way we got a sufficiently large corpus of un-parallel texts for Latvian and Lithuanian.

For the experiment we selected 27 Lithuanian terms and 80 corresponding Latvian term candidates. Only terms with at least 50 occurrences in corpus were selected and only Lithuanian terms for which there were at least one correct and one incorrect Latvian term were selected. Correct translation was depicted by human terminologist.

Every Lithuanian term had from 2 to 8 candidate translations in Latvian from which only 1 to 4 were correct.

We implemented the proposed method and experimented with different parameter p threshold settings.

The size of window for collocations was 10 words to the left and right of the term occurrences. As Latvian and Lithuanian are highly inflected languages, morphological normalization was applied.

To measure the usefulness of our method we chose the value of threshold parameter p = 0.002.

We say that our method gives correct result on Latvian term *s* (which is translation candidate of Lithuanian term *t*):

- If *s* is a correct translation of *t* and its score is at least 5% higher than for every other incorrect translation candidate of *t*;~~.~~
- If *s* is not a correct translation of *t* and its score is at least 5% lower than for any other correct translation of *t*;~~.~~
- If scores of correct translation and an incorrect translation of *t* differ by less than 5% then we assume that the difference in score is not sufficiently significant[IS14];
- Otherwise we say that our method returns an incorrect [IS15]result.

Examples of results are in Figure 1 and Figure 2. On X axis there are different values of threshold p and on Y axis are the scores for term pair.

Results of experiment showed that our method returned a correct answer in 61% of cases (for 49 out of 80 Latvian terms).

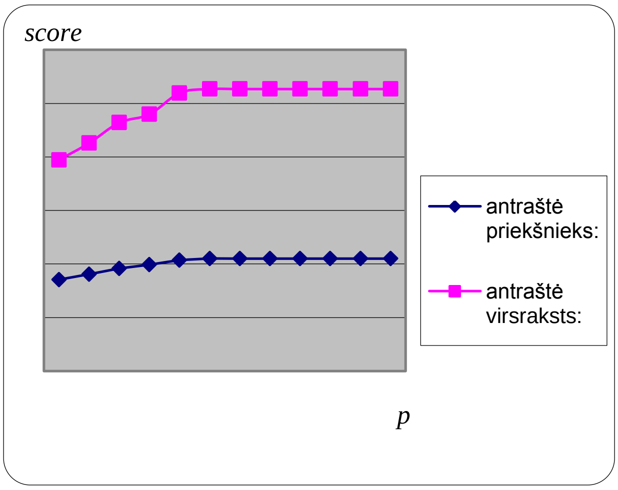


Figure 13 Correct Latvian term virsraksts for Lithuanian term antraštė achieved significantly higher score than wrong translation priekšnieks

For 21% (17 out of 80 Latvian terms) difference in score[IS16] was no sufficiently significant.

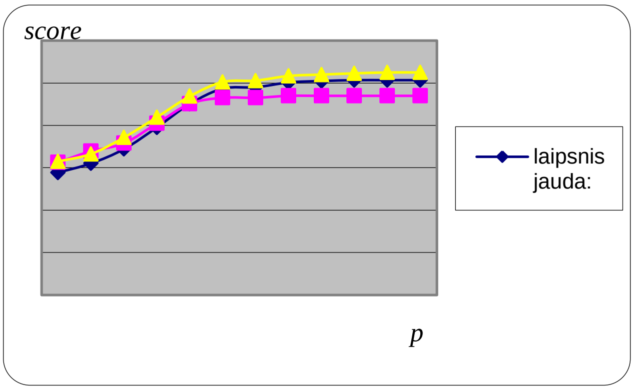


Figure 14 Example of insufficient difference: [IS17]score for Latvian term jauda and Lithuanian term laipsnis.

For 18% (14 out of 80 Latvian terms) the method gave wrong [IS18]result.

6.2 INTEGRATION IN AUTHORING SYSTEMS

Typically, translators spend 30-60% of total translation time on terminology research. Therefore, it is of vital importance to ensure that they can use all the required terminology resources in the right format and in a convenient environment. Increasingly, terminology research is done using sources that are available on the Internet. Currently, translators spend a lot of time inefficiently, searching and processing information from multiple online sources, copy-pasting or changing the format to the one that they require in their work environment. Spending time on technical aspects instead of focusing on true terminology research results in cost inefficiencies and reduced translation quality.

Faced with difficulties in accessing the terms they need and participating in collaborative activities to create new terms, many translators create their own terminology resources. They typically store these terms in a spreadsheet or other proprietary formats that are not efficiently connected to a multitude of translation environments that they might use. Moreover, these resources are not shared with other translators and potential users. This results in redundant work or even reduced translation quality and does not bring additional value to the creator of this custom terminology.

To increase efficiency and quality of translation, translators need an easy access to multiple terminology databases, facilities to enable collaborative efforts in creation of new terms, productivity tools to get necessary terms right from translation environment (Lengyel and Vasiljevs, 2008).

There have been several efforts to provide reasonable solutions to support translators accessing multilingual terminology resources. For example, Quest tool brings consolidated terminology content closer to its user and is used internally by translators in the DG for Translation of the European Commission (European Commission Directorate-General for Translation, 2008).

Although our terminology consolidation methodology provides single access point to variety of terms, still an extra effort required from the user to switch from translation environment to terminology web-page, specify the search query, select a result and go back to translation tool and type the term there.

We propose solution were access to online terminology databases is supported directly from the most widely used translation environments, such as SDL Trados, Deja Vu, Wordfast, MemoQ, as well as other applications that are commonly used in the translation process, such as Microsoft Word, PowerPoint, Excel as well as open-source applications.

Such solution includes:

- Terminology integration component for instant access from the text editing environment to the web-based terminology data by invoking web-service based queries;
- External termbase API to enable enables third party software manufactures to provide their users with direct access to the content of termbase.

This is especially useful in the translation usage scenario since such a solution will deliver well-targeted content from termbase to productivity environments used routinely by translators and other language workers.

6.2.1 USAGE SCENARIOS

Target clients of terminology integration component can be segmented as follows:

- Translation service providers:
    - Freelance translators (and other individual users: editors, technical writers, etc.);
    - Translation agencies;
    - Localization service providers.
- Translation service consumers (using outsourced and / or in-house services):
    - Commercial companies with products / services in global markets;
    - International organizations;
    - EU institutions;
    - National government institutions.
- Minor client groups:
    - Various organizations: term banks, libraries, publishing houses; universities, providers of web-based CAT tools etc.;
    - Various individuals: students, academia, "general reference" users, etc.

Nevertheless, freelance translators and in-house translators are foreseen to be major target user groups for the new tool.

A typical portrait of a freelance translator is as follows:

- Produces translations for various clients;
- Works from a home office;
- May use a CAT tool like Trados or Wordfast, but not necessarily;
- Usually works from 1-2-3 source languages into 1-2 target languages;
- Often specializes in a certain subject area;
- Price elasticity is high - has to finance IT purchases from own budget;
- IT literacy varies from very high to moderate and low.

A typical portrait of an in-house translator is as follows:

- Produces translations for a certain employer which is often a translation agency / localization service provider / government institution / commercial company;
- May use a CAT tool like Trados or Wordfast, but not necessarily;
- Usually works from 1-2-3 source languages into 1-2 target languages;
- Often specializes in a certain subject area;

- Is less price-sensitive, as IT purchases are done by employer;
- IT literacy varies from very high to moderate and, in rare cases, low.

To determine target user groups' expectations regarding terminology integration component and use of termbase resources beyond the portal, a survey of user needs was carried out. A questionnaire was developed and sent via e-mail to 80 translators. Among questions were inquiries about the platform, software, Internet resources and search engines used in translation process as well as inquiries about workflow organization and terminology management.

In fact, the survey results show that translators and terminologists have difficulty envisioning new computer software tools, and they prefer using existing tools they have gotten used to already. The quantitative results of the survey are as follows:

80% respondents use Microsoft Word in their everyday workflow, 25% use SDL Trados in conjunction with Microsoft Word, the rest use free tools like OpenOffice or Globalsight;

80% respondents do not manage terminology, the rest use MultiTerm or SDL Trados;

40% respondents process files in Microsoft Word format, 25% – in html, 15% – in pdf and 20% in a mix of other formats (e.g. Excel or PowerPoint formats).

Furthermore, about 90% respondents use Google for terminology research. Nevertheless, the survey results show users' interest and necessity for additional terminology tools especially for Microsoft Office (with high priority for supporting Microsoft Word, lower – for Microsoft PowerPoint and OpenOffice Writer, and the lowest – for Microsoft Excel and e-mail writing tools). Therefore, the biggest potential for meeting users' needs is in supporting Microsoft Office. Besides, Microsoft Word integrates with SDL Trados and thus bridges the gap to the user of CAT tools. Both Microsoft Word 2003 and 2007 should be supported since the first is still popular among many users and the second offers broader functionality.

The goal is to provide an access to online terminology content with a single keyboard shortcut, even without opening a browser window.

Conceptual design of the integration component should comprise the following components (Figure 1):

Termbase should provide users with the following key functions (see the description below):

- Search;
- Segment analysis;
- Entry request.

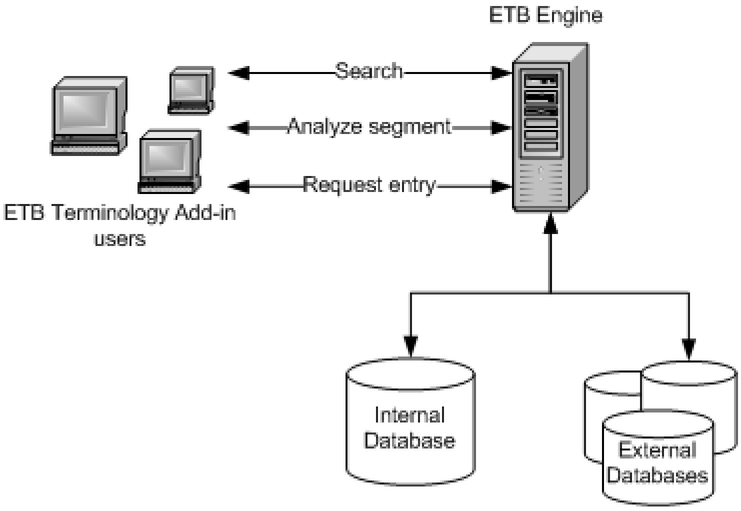


Figure 15 Design of EuroTermBank Terminology Add-in.

To evaluate the proposed solution, EuroTermBank Terminology Add-in was developed and evaluated.

It meets the following requirements:

- Easy download, quick setup, low usage of computer resources;
- Integration into Word Research pane and compact / clear arrangement of terms in it;
- Intuitive use of the tool and no hidden or complicated features, keyboard shortcuts.

The EuroTermBank Terminology Add-in integrates into Microsoft Word as follows:

- A button "EuroTermBank" in Microsoft Word in the Review Ribbon, Terminology Group (Figure 16);
- A custom Microsoft Research pane "EuroTermBank Terminology" (Figure 17);
- Contextual menu when right-clicking a text segment (Figure 18).

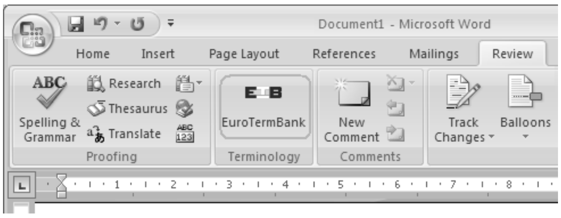
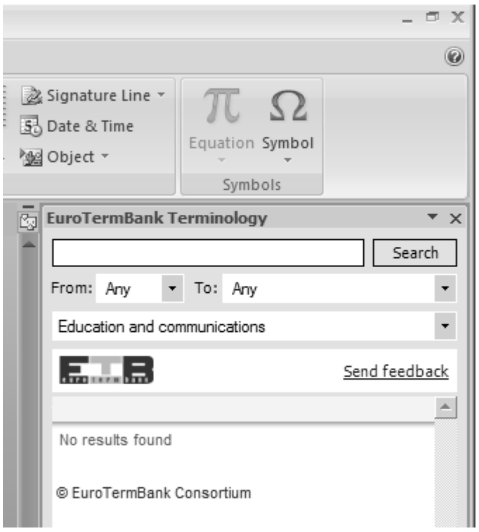


Figure 16 EuroTermBank button in the MS Word ribbon.



From: Any To: Any

Education and communications

No results found

© EuroTermBank Consortium

Figure 17 Terminology access pane.

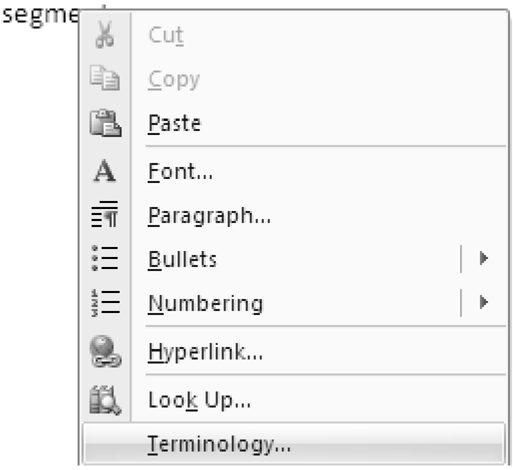
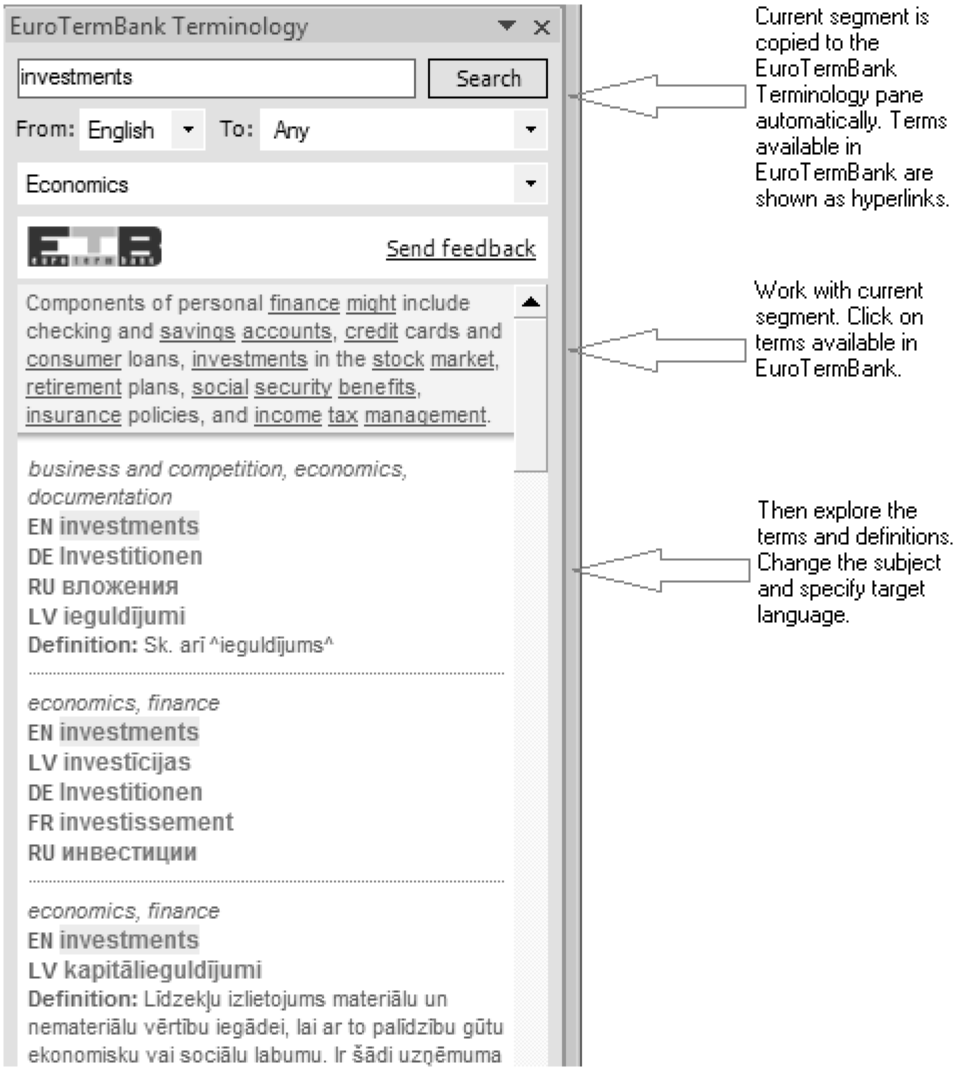


Figure 18 Invoking terminology access through contextual menu.

The EuroTermBank Terminology Add-in provides the following functionality:

- Provides targeted search results in Word Research pane;
- Filters terminology by domain;
- Filters terminology by language;
- Automatically detects source language;
- Identifies terms in a segment / sentence and researches the EuroTermBank internal and external resources for the identified terms;
- Provides the option of reaching full search results by opening the EuroTermBank portal in a web browser;
- Provides user feedback function.

It should be mentioned that the function of identifying terms in a segment or sentence (Figure 19) and then searching the EuroTermBank resources for them is highly appreciated by end users. The tool identifies terms and shows them hyperlinked in the topmost part of the pane. Moreover, the user can change the language and domain settings, and the tool updates the relevant links in specified languages or domains.



investments

From: English To: Any

Economics

Components of personal finance might include checking and savings accounts, credit cards and consumer loans, investments in the stock market, retirement plans, social security benefits, insurance policies, and income tax management.

business and competition, economics, documentation
EN investments
DE Investitionen
RU вложения
LV ieguldījumi
Definition: Sk. arī ^ieguldījums^

economics, finance
EN investments
LV investīcijas
DE Investitionen
FR investissement
RU инвестиции

economics, finance
EN investments
LV kapitālieguldījumi
Definition: Līdzekļu izlietojums materiālu un nemateriālu vērtību iegādei, lai ar to palīdzību gūtu ekonomisku vai sociālu labumu. Ir šādi uzņēmuma

Current segment is copied to the EuroTermBank Terminology pane automatically. Terms available in EuroTermBank are shown as hyperlinks.

Work with current segment. Click on terms available in EuroTermBank.

Then explore the terms and definitions. Change the subject and specify target language.

Figure 19 Identification of terms in a sentence.

Beta versions of EuroTermBank Terminology Add-in for Microsoft Word 2003 and 2007 can be freely downloaded from the EuroTermBank multilingual terminology portal.

6.2.2 USER FEEDBACK

The developed tool was tested and evaluated by end users before its release (internal beta testing) and after it (external testing). The following two subsections provide the description of user feedback.

The developed tool was tested by localization specialists for two month before release. 20 full-time and about 50 part-time professional translators were involved in beta testing. To receive feedback from terminologists, they were asked to fill in a questionnaire. Altogether 80 out of 148 sent questionnaires were filled in.

General results of the internal beta testing show that:

- 30% respondents (more than a half) consider the developed tool to be extremely useful and efficient for their translation needs;
- At the same time 40% respondents define the tool as good software which could be of help for their work;
- 70% respondents point out that an access to the terms is simple and fast (contextual menu, in particular);
- Nearly all respondents appreciate language coverage.

The main suggestions are as follows:

- 10% respondents suggest that instructions about how to use the tool should be built in as an internal function;
- 20% respondents find that they would like to have a localized web page and instructions;
- 89% respondents point out that the amount of displayed search results is too large and same results appear to be from different sources.

After the release of beta version we also got users' feedback. Among advantages users emphasize user-friendly interface and segment / sentence analysis function.

Users point out the following functionality to be developed:

- Built-in help;
- Macintosh support;
- Short keys change;
- Saving of user settings (e.g. translation direction).

Users also suggest adding new languages, for example, Persian.

6.3 TERMINOLOGY DATA SHARING

Sharing of literally everything that someone finds interesting, amusing or valuable is a true Web 2.0 phenomenon, and the key concept in sharing is voluntary user participation. In a specific way, this phenomenon can be identified in the area of terminology resources as well. Abovementioned survey shows that 37% or terminology users are willing to share their resources.

Terminology sharing typically involves sharing of non-confidential, non-competing and non-differentiating terminology across various actors – individuals along with companies and language service providers, often with the goal to consolidate and promote accessibility to multilingual terminology per vertical industries (Rirdance, 2007). Terminology sharing involves returns from streamlined industry terminology, by ensuring reuse of existing terminology assets. For those who share their terminology, it is a way of promoting and disseminating one's well-established terminology, possibly even to the level of *de facto* industry standard terminology.

Industry players have the following key benefits from terminology sharing:

- Promotes establishment of an industry standard terminology;
- Helps to develop and enhance industry terminology, particularly for the minor languages, in a cost-efficient way, resulting in improved quality and user experience for localized products;
- Stimulates harmonization and unification of industry terminology, usage of common terms for common concepts across different products and vendors, enhancing overall user experience and shorter learning curve;
- Helps in transition towards more open and cost-efficient translation and localization business models, reducing the overhead of intermediary suppliers with little or no value added (or, sometimes, with value reduced);
- Distinguishes vendor specific terms – terms that are associated with particular features and concepts differentiating vendor's products from the products of the competition;
- Highlights vendor's contribution to greater community values such as national languages which are key and the most sensitive elements of national identity especially in smaller countries;
- Enhances public availability of language resources thus supporting the research and development of language technologies, particularly for minor languages;

- Strengthens vendor’s market position by boosting user involvement in the particular brand and products, and nurturing growth of communities around particular products.

Terminology sharing on EuroTermBank provides several additional benefits:

- Increases the dissemination of vendor’s terminology through the largest on-line terminology data base to professional communities and a large user base, across the European Union marketplace;
- Adds vendor’s terminology to the already respected and reliable multilingual terminology sources of EuroTermBank such as national terminology databases and IATE – inter-institutional terminology database of the European Union;
- Provides direct and easy access to vendor’s terms to the professional translation community through EuroTermBank professional access tools for Microsoft Word, SDL Trados and other desktop authoring environments;
- Facilitates convergence of terminology used in practice and in official documents as EuroTermBank is one of the sources for terminology search in institutions of European Union and member countries;
- Supports machine translation development for minor languages and narrow domains as EuroTermBank resources are used in European Union research projects in machine translation.

These there some of the reasons why Microsoft selected EuroTermBank as a data sharing platform for their multilingual terminology data. According to a survey (Gornostay, 2010) Microsoft Language Portal is the third most used online terminology portal. Microsoft is among pioneers in industry data sharing on public online repositories expanding EuroTermBank with more than 20 000 information and communication technology terms in 26 languages.

Significant development in the area of sharing linguistic resources is also TAUS Data Association that positions itself as “a super cloud for the global translation industry, helping to improve translation quality, automation and fuel business innovation”[9]. Although mostly oriented towards sharing translation memories, it does involve sharing of terminology resources as well.

[9] http://www.tausdata.org

To reap the full benefits from the shared terminology, it is essential to ensure integrated access to these terminology resources in translation environments.

However, the concept of sharing is not really present in major terminology banks. Instead of providing the opportunity for users to contribute their own resources or share their findings over social networks, terminology banks typically keep to the traditional one-way communication of their high-quality preselected resources.

6.4 TARGETED DELIVERY

A further step in the direction of meeting user expectations and providing the required terminology resources to its users in a most efficient way involves integration of content delivery in the production environments of terminology users.

To deliver targeted content from the EuroTermBank portal to its users, a layer of connectivity tools is being developed for terminology research in specific work environments, such as plug-ins for use with Microsoft Word (released), SDL Trados and MemoQ (upcoming) (Gornostay et al, 2010). SDL Trados being the most popular tool of choice for professional translators and Microsoft Word being used by general purpose users as well as translation professionals, EuroTermBank content is accessible to majority of its users with a single keyboard shortcut, without opening a browser window. The Microsoft Word plug-in provides the following functionality:

- Filters terminology by subject;
- Filters terminology by language;
- Automatically detects source language;
- Identifies terms in a sentence/segment;
- Provides targeted search results in Microsoft Research pane;
- Provides the option of reaching full search results by opening the portal in a web browser.

Quest is a similar tool that brings consolidated terminology content closer to its user. This metasearch interface which translators can use to query several databases simultaneously is used internally by translators in the Directorate-General for Translation of the European Commission and was developed „with a view to centralizing, simplifying and speeding up terminology searches" (EC DGT, 2008). A

Quest search can be launched by pressing a button in Microsoft Word; translators can select the source and target language pair, and one of three available profiles determining which databases they wish to search. However, this tool is not made available to the general public.

Of course, connectivity could also be provided and supported from the side of translation tools. Although a number of translation tools already provide basic integration with terminology web searches, e.g., the user can define a number of term banks to be queried, the nature of these features is such that they will necessarily be general and not adapted to specifics of each term bank, thus possibly making the results of these searches quite useless.

6.5 USER PARTICIPATION

The new paradigm of using World Wide Web resources supports active user involvement in shaping and elaborating the content of web resources. In respect to terminology resources, this area, however, seems to have been lagging behind introduction of other Web 2.0 concepts. Successes of efforts in encouraging user participation in public terminology forums can be described as limited for a number of reasons: participants to terminology forums are split by language; the audience interested in active terminology discussions is limited; keepers of online terminology banks are concerned about maintaining high quality standards, hence they do not provide opportunities to actively shape terminology content, as opposed to welcoming feedback and comments.

However, this means that terminology content in term banks does not benefit from the content that could be provided by their users, or from the quality improvements that could be implemented by these users. Encouraging true user participation and reaping benefits from it remains the biggest challenge for term banks in adopting the new approaches collectively referred to as Web 2.0.

A potential road towards encouraging true user participation would include removing the current obstacles for participation. Among other things, this would involve opening up term banks for sharing of user terminology; creation of a staged validation or voting system. With professional translators comprising the largest user group of term banks, term banks should support instant “capturing” of a term or a term candidate, so that users could submit terms with a single click from their productivity environments.

7. CONCLUSIONS

This research is the first work dedicated to an integrated view on the problem of consolidation of heterogeneous multilingual terminology resources. Both theoretical and practical guidelines are provided covering major aspects in consolidation and representation of terminological data.

The analysis and conclusions are based on extensive studies of the best practice in the field and an evaluation of international standards for their applicability and adaptation in real life scenarios. This greatly facilitates adaptation and implementation of standards and ensuring interoperability of global terminology resources.

Requirements analysis method is proposed based on terminology work scenarios. Three distinctive scenarios are identified from the perspective of goals and conditions of terminology work – local, national and international scenarios.

Terminology data requirements are analyzed using scenario based view and data modeling principles are proposed. The necessity to use concept-oriented modeling principles is substantiated and shortcomings of lexicographical modeling are demonstrated. Data modeling for international scenario is elaborated adopting a four layer data structure – entry level, language level, term level and word level.

Data storage and exchange standards are analyzed and optimal formats are recommended. TBX standard is recommended for data exchange. Applicability of TBX for data storage is proposed and demonstrated experimentally.

The federation principle is proposed for consolidation of independently maintained terminology databases. This solves a problem of consolidation of independent heterogeneous termbases. The federated approach in consolidation of resources enables distributed terminology to be accessible through a central gateway while it is maintained locally.

An entry compounding mechanism is introduced for unified representation of terminology data. Entry compounding greatly facilitates consolidation of national terminology resources into a multilingual system. Basic compounding mechanisms are practically implemented and demonstrated. Further improvements using corpus based analysis are suggested and experimentally affirmed.

A practical demonstration of the research results in EuroTermBank project serves as a proof-of-concept for the proposed approaches. The largest online source of terminology in multiple subject fields in languages of new European Union countries is developed with about 2 mil. term entries.

As a result of this work, a new type of international terminology infrastructure is proposed and implemented providing access to diverse terminology resources and providing basis for further consolidation of terminology in Europe and beyond. It can facilitate research and practical work in terminology, lexicography, computational linguistics, as well as applied in computer-assisted translation systems.

We can conclude that results of the thesis work prove the research hypothesis and access and usability problems posed by fragmentation and heterogeneity of terminology resources can indeed be effectively solved with a federated multilingual terminology portal that provides consolidated data representation and integration in authoring software.[IS19]

BIBLIOGRAPHY

AUTHOR'S PUBLICATIONS

Vasiljevs, A., Rirdance, S. & Gornostay, T., 2010. Reaching the User: Targeted Delivery of Federated Content in Multilingual Term Bank. In *Proceedings of the TKE (Terminology and Knowledge Engineering) Conference 2010*. Dublin, 2010.

Vasiļjevs, A. & Balodis, K., 2010. Corpus based analysis for multilingual terminology entry compounding. In *Proceedings of LREC 2010 Conference*. Malta, 2010.

Vasiljevs, A., Rirdance, S. & Balkanyi, L., 2008. Ontological Enrichment of Multilingual Terminology Databank. In *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering TKE 2008*. Copenhagen, 2008, pp.279-289.

Vasiljevs, A. & Rirdance, S., 2008. Application of terminology standards for a multilingual term bank: the EuroTermBank experience. In *Proceedings of the LREC-2008 Workshop on Uses and usage of language resource-related standards*. Marrakech, 2008.

Vasiljevs, A., Rirdance, S. & Liedskalnins, A., 2008. EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources* ICGL 2008. Hong Kong, 2008, pp.213-220.

Vasiljevs, A. & Rirdance, S., 2007. Towards Consolidation of Dispersed Multilingual Terminology Resources. *International Journal of Translation, Special Issue on Lexical Resources and Machine Translation*, 19(1), Bahri Publications, New Delhi, 2007, ISSN 0940-9819, pp.65-77.

Liedskalniņš, A., Vasiļjevs, A. & Rirdance, S., 2007. From Paper to TBX: Processing Diverse Data Formats for Multilingual Term Bank, Human Language Technologies. In *Proceedings of the Third Baltic Conference "Human Language Technologies – the Baltic Perspective"*. Kaunas, 2007.

Vasiljevs, A. & Rirdance, S., 2007. Consolidation and unification of dispersed multilingual terminology data. In *International Conference RANLP 2007 (Recent Advances in Natural Language Processing)*. Borovets, 2007, pp.614-618.

Henriksen, L., Povlsen, C. & Vasiljevs, A., 2006. EuroTermBank – a Terminology Resource based on Best Practice. In *Proceedings of the LREC 2006, the 5th International Conference on Language Resources and Evaluation*. Genoa, 2006.

Vasiļjevs, A., Borzovs, J., Skadiņš, R. & Liedskalniņš, A. 2006. Development of web-based terminology database for new EU member countries – problems and opportunities. In *Proceeding of the Seventh International Baltic Conference on Databases and Information Systems Baltic DB&IS 2006*. Vilnius, 2006, pp.228-238.

Skujiņa, V., Ilziņa, I., Vasiļjevs, A. & Borzovs, J. 2006. Terminology Standards in the Aspect of Harmonization for International Term Database. *Terminologija*, 13, Lietuvių kalbos institutas, Vilnius, 2006, pp.17-32.

Vasiljevs, A. & Schmitz, K.-D. 2006. Collection, harmonization and dissemination of dispersed multilingual terminology resources in an online terminology databank. In *Proceedings of TSTT 2006, Third International Conference on Terminology, Standardization and Technology Transfer*. Beijing: Encyclopedia of China Publishing House, 2006, pp.265-272.

Skadiņš, R. & Vasiļjevs, A. 2004. Multilingual Terminology Portal – termini.letonika.lv. In *Proceedings of the First Baltic Conference ”Human Language Technologies – the Baltic Perspective”*. Riga, 2004, pp.183-186.

Gornostay, T., Vasiljevs, A., Rirdance, S. & Rozis, R., 2010. Bridging the Gap – EuroTermBank Terminology Delivered to Users’ Environment. In *Proceedings of 14th Annual Conference of the European Association for Machine Translation*. Saint-Raphaël, 2010.

OTHER PUBLICATIONS

Ahmad, K. et al., 1996. *POINTER Final Report*. [Online] Available at: http://www.computing.surrey.ac.uk/ai/pointer/report/index.html [Accessed 19 June 2010].

Anon., 2009. The ISO Concept Database ISO/CDB released New tool for standards development and use of standards. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement*.

Auksoriūtė, A., Gaivenytė, J. & Umbrasas, A., 2003. The state of Lithuanian terminology. *Terminoloģijas jaunumi*.

Balodis, K., 2010. *Teksta korpusu datoranalīze dažādu valodu terminu ekvivalences noteikšanai*. Bakalaura darbs. Rīga: Latvijas Universitātes Datorikas fakultāte.

Betz, A. & Schmitz, K.-D., 1999. The Terminology Documentation Interchange Format TeDIF. In *Terminology and Knowledge Engineering TKE'99*. Innsbruck, Wien, 1999.

Beyer, H. & Holtzblatt, K., 1998. *Contextual Design: Defining Customer-Centered Systems*. London: Academic Press.

Blaschke, C., 2003. Distributed Terminology Management: Modern Technologies in Client/Server Environments. In *Proceedings of the 6th International TAMA Conference*. Pretoria, 2003.

Boguslavsky, I., Cardeñosa, J. & Carolina, G., 2009. A Novel Approach to Creating Disambiguated Multilingual Dictionaries. *Applied Linguistics*, 30(1), pp.70-92.

Boguslavsky, I. & Dikonov, V., 2008. Universal Dictionary of Concepts. In *Proceedings of MONDILEX First Open Workshop*. Moscow, 2008.

Brinkmann, K.-H., 1980. Terminology data banks as a basis for high-quality translation. In *Proceedings of the 8th conference on Computational linguistics*. Tokyo, 1980.

Cabré, T.M., 1999. *Terminology: theory, methods and applications*. Philadelphia: John Benjamins Publishing Company.

Cabré, T.M., 2003. Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2), pp.163-99.

Chiocchetti, E. & Voltmer, L., eds., 2008. *Harmonising Legal Terminology*. Bolzano/Bozen: EURAC.

COTSOES, 2002. ISBN 3-907871-07-3 *Recommendations for Terminology Work*. Berne: MediaCenter of the Confederation.

DCMI, 2007. *DCMI Abstract Model*. [Online] Available at: http://www.dublincore.org/documents/abstract-model/ [Accessed 18 June 2010].

Depecker, L., 2010. *Terminologie*. [Online] Available at: http://www.universalis.fr/encyclopedie/terminologie/ [Accessed 20 June 2010].

DIN, 2010. *The DIN Terminology Database*. [Online] Available at: http://www.beuth.de [Accessed 20 June 2010].

Dubuc, R., 1972. TERMIUM System Description. *Translators' Journal*, 17(4), pp.203-19.

Esselink, B., 2000. *A Practical Guide to Localization*. Amsterdam/Philadelphia: John Benjamins Publishing.

EuroTermBank, 2005. *Current standards and best practices assessment report*. [Online] Available at: http://project.eurotermbank.com/uploads [Accessed 20 June 2010].

Faber, P., Leon, P. & Prieto, J.A., 2009. Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies*, 2009(1).

Faber, P., Márquez Linares, C. & Vega Expósito, M., 2005. Framing Terminology: A Process-Oriented Approach. *Meta: Translators' Journal*, 50(4).

Felber, H., 1984. *Terminology manual*. Paris: Unesco: International Information Centre for Terminology (Infoterm).

Galinski, C., 2005. Semantic Interoperability and Language Resources. In *Terminology and Content Development. – Copenhagen*. Copenhagen, 2005.

Galinski, C., 2007. New ideas on how to support terminology standardization projects. *eDITion*, 1.

Gorjanc, V., Krek, S. & Vintar, Š., 2008. Slovene Terminology Web Portal. In *Proceeding of Euralex 2008*. Barcelona, 2008.

Gornostay, T., 2010. Terminology management in real use. In *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*. Saint-Petersburg, 2010.

Harold, E.R., 2005. [Online] Available at: http://www-128.ibm.com/developerworks/xml/library/x-mxd1.html [Accessed 20 September 2007].

Hodge, G., 2000. *PowerPoint presetation "Federating Terminology: Can We Avoid Reinventing the Wheel?"* [Online] Available at: http://www.chin.gc.ca/Resources/Cidoc/English/Presentations/ghodge.html [Accessed 20 June 2010].

Infoterm, 2005. *Guidlines for Terminology Policies*. Paris: UNESCO. http://unesdoc.unesco.org/images/0014/001407/140765e.pdf.

ISO 10241, 1992. *ISO 10241: International terminology standards -- Preparation and layout*. Geneva: ISO.

ISO 1087-1, 2000. ISO 1087-1:2000 *Terminology work — Vocabulary — Part 1: Theory and application"*. Geneva: ISO.

ISO 22128, 2008. *ISO 22128:2008 Terminology products and services - Overview and guidance*. Geneva: ISO.

ISO 26162, 2010. *ISO 26162:2010 Systems to manage terminology, knowledge and content - Design, implementation and maintenance of Terminology Management Systems (Draft International Standard)*. Geneva: ISO.

ISO 5964, 1985. *ISO 5964:1985 Documentation - Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO.

ISO 704, 2009. *ISO 704:2009 Terminology work — Principles and methods*. Geneva.

ISO, 1.-1., 2000. *Terminology work — Vocabulary — Part 1: Theory and application"*. Geneva: ISO.

IZM, Latvijas Republikas Izglītības un zinātnes ministrija, 2006. *Valsts valodas politikas programma 2006.-2010. gadam*. Rīga: Ministru kabinets.

Johnson, I. & Macphail, A., 2000. IATE - Inter-Agency Terminology Exchange: Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. In *Proceedings of the Workshop on Terminology Resources and Computation 2000, LREC 2000*. Athens, 2000.

Kashyap, V., 2000. *Information Brokering Across Heterogeneous Digital Data: A Metadata-based Approach*. Kluwer Academic Publishers.

Kent, A., ed., 1998. *Encyclopedia of library and information science*. New York: CRC Press.

Laurent, J., 1977. Utilization of the Technical Terminology Standardized at AFNOR. In *In: Overcoming the language barrier*. Munich, 1977. Verlag Dokumentation.

Liedskalniņš, A., 2007. *Daudzvalodu terminoloģisko datu glabāšana EuroTermBank sistēmā, Maģistra darbs*. Latvijas Universitāte.

Lyding, V., Chiocchetti, E., Sérasset, G. & Brunet-Manquat, F., 2006. The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Sydney, 2006. Association for Computational Linguistics.

Lyding, V., Chiocchetti, E., Sérasset, G. & Brunet-Manquat, F., 2006. The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated. In *Proceedings of the workshop on multilingual language resources and interoperability*. Sydney, 2006. Association of Computational Linguistics.

McNaught, J., 1993. Terminological Data Banks: a model for a British Linguistic Data Bank (LDB). In *Proceedings of ASLIB.*, 1993.

Musser, J. & O'Reilly, T., 2006. *Web 2.0 Principles and Best Practices*. O'Reilly Radar.

Nilsson, H., 2010. Towards a national terminology infrastructure. The Swedish experience. In M. Thelen & F. Steurs, eds. *Terminology in Everyday Life*. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp.61-80.

Nkwenti-Azeh, B., 1993. New trends in terminology processing and implications for practical translation. In *Proceedings of ASLIB.*, 1993.

Nuopponen, A., 1996. Terminological information and activities in World Wide Web. In *Proceedings of TKE'96, Terminology and Knowledge Engineering*. Frankfurt, 1996. INDEKS-Verlag.

Pavel, S. & Nolet, D., 2001. *Handbook of Terminology*. Quebec: Translation Bureau, Public Works and Government Serices Canada.

Pohn, R. & Weissinger, R., 2008. Zooming in on the ISO Concept database. *Terminology Standardization and Harmonization (TSH)*, 35/36, pp.7-10.

Prószéky, G., 1998. An intelligent multi-dictionary environment. In *Proceedings of the 17th international conference on Computational linguistics*. Montreal, 1998. ACL.

Qian, D. & Teng, X., 2009. *Localization and Translation Technology in the Chinese Context*. [Online] Available at: http://tac-online.org.cn/en/tran/2009-10/13/content_3183433.htm [Accessed 20 June 2010].

Rirdance, S. & Vasiljevs, A., eds., 2006. *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project*. Riga: EurotermBank Consortium.

Rummel, D. & Ball, S., 2001. The IATE Project – Towards a Single Terminology Database for the EU. In *Proceedings of ASLIB 2001, the 23rd International Conference on Translation and the Computer*. London, 2001.

Sager, J.C., 1990. *A Practical Course in Terminology Processing*. Amsterdam & Philadelphia: John Benjamins.

Sager, J.C. & McNaught, J., 1980. Feasibility study of the establishment of a terminological data bank in the U.K. *UNESCO Alsed-LSP Newsletter*.

Schulz, J., 1980. A Terminology Data Bank for Translators (TEAM). *META: Translators' Journal*, 25(2), pp.211-29.

Seimas, L., 2003. *Lietuvos Respublikos Terminų Banko Įstatymas*. Vilnius: Lietuvos Respublikos Seimas.

Sérasset, G., Brunet-Manquat, F. & Chiocchetti, E., 2006. Multilingual Legal Terminology on the Jibiki Platform: The LexALP Project. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. Sydney, 2006.

Sheth, A.P. & Larson, J.A., 1990. Federated database systems for managing distributed, heterogeneous, and autonomous database. *ACM Computing Surveys*.

Somers, H., 2003. *Computers and Translation. A translator's guide.* John Benjamins Publishing Company.

Steinberger, R. et al., 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *5th Intl. Conf. on Language Resources and Evaluations*. Genoa, 2006.

Streiter, O. & Voltmer, L., 2003. A Model for Dynamic Term Presentation. In *Proceedings of TIA-2003 Conference*. Strasbourg, 2003.

Streiter, O., Voltmer, L., Ties, I. & Lyding, V., 2005. Structuring Terminological Data: The BISTRO Proposal. *Terminology Science and Research*, 16.

Thelen, M. & Steurs, F., eds., 2010. *Terminology in Everyday Life*. Amsterdam/Philadelphia: John benjamins B.V.

Thurmair, G., 2006. Exchange Formats: TBX, OLIF and beyond. *LDV-Forum*, 21(1), pp.45-56.

UNESCO, 2005. *Guidelines for Terminology Policies. Formulating and implementing terminology policy in language communities / Prepared by Infoterm*. Paris: UNESCO.

Vasiļjevs, A., 2008. The influence of new technologies upon the Latvian language. In *Break-out of Latvian*. Rīga: Zinātne. pp.345-55.

Vasiļjevs, A. & Rirdance, S., 2008. Latviešu valodas terminu konsolidēšana vienotā terminu bankā. In *Letonikas otrais kongress. Valodniecības raksti-2.*. Rīga, 2008. Latvijas Zinātņu akadēmija.

Vasiļjevs, A. & Skadiņš, R., 2004. Multilingual Terminology Portal – termini.letonika.lv. In *The First Baltic Conference "Human Language Technologies – the Baltic Perspective"*. Riga, 2004.

Warburton, K., 2007. *Standards and Guidelines for the Language Industry*. [Online] Language Technologies Research Centre: Language Technologies Research Centre (2) Available at: http://www.crtl.ca/docs/StandardsAndGuidelinesForTheLangIndustr_FINAL_July_2009.pdf [Accessed 19 June 2010].

Weissinger, R., 2007. Standards as databases and the development of knowledge. *ISO Focus*, November 2007. Online publication (http://www.infoterm.info/pdf/activities/TSH/TSH33.pdf).

Weissinger, R., 2008. *ISO Concept Database presentation*. [Online] Available at: http://www.iso.org/iso/livelinkgetfile?llNodeId=148795&llVolId=-2000 .

Wright, S.E., 2007. Coping with indeterminacy. Terminology and knowledge representation resources in digital environments. In B.E. Antia, ed. *Indeterminacy in Terminology and LSP*. Amsterdam/Philadelphia: John Benjamins B.V. pp.157-80.

Wright, S.E. & Budin, G., 1997. *Handbook of Terminology Management*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Wright, S.E. & Budin, G., 2001. *Handbook of Terminology Management*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Wüster, E., 1968. *The Machine Tool: An Interlingual Dictionary of Basic Concepts*. London: Technical Press.

Wüster, E., 1972. *Einführung in die Allgemeine Terminologielehre und terminologische Lexikographie*. Vienna: Infoterm.

LIST OF ABBREVIATIONS

API	Application Programming Interface
ClaML	Classification Markup Language
ECDC CTS	Core Terminology Server of the European Centre for Disease Prevention and Control
EU	European Union
HLT	Human Language Technologies
HTTP	Hypertext Transfer Protocol
ISO	International Organization for Standardization
ISO/CDB	ISO Concept database
LGP	Language for General Purpose
LSP	Language for Special Purpose
LZA TK	Terminology Commission of the Academy of Sciences of Latvia
OWL	Web Ontology Language
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TMS	Terminology management system
W3C	World Wide Web Consortium
WHO	World Health Organization
XML	Extensible Markup Language

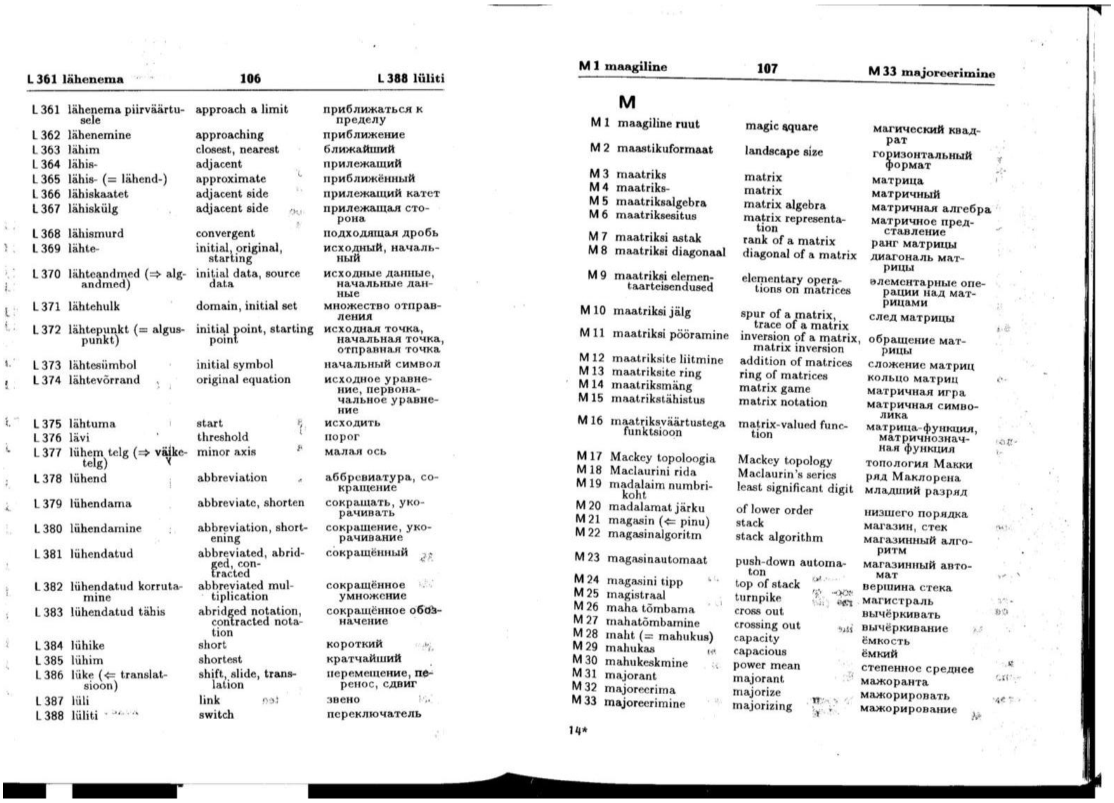
APPENDIXES

APPENDIX 1. DATA CATEGORIES OF EUROPEAN INTER-INSTITUTIONAL TERMBASE IATE

Levels	IATEdata fields	
Language independent level	LIL_RECORD	CREATION_DATE
		CHANGED_BY
	INSTITUTION	CHANGE_DATE
	AUTHOR	CHANGED_IN_FIELDS
	PROPOSER	
	MARKED_FOR_DELETION_MERGING	DOMAIN
	CONFIDENTIALITY	DOMAIN_NOTE
	DATE_MADE_CONF	ORIGIN
	MADE_CONF_BY_USER	ORIGIN_NOTE
		PROBLEM_LANG_CODE
		COLLECTION
		CROSS_REFERENCE
		GRAPHICS
Language level	LIL_RECORD	LANGUAGE_USAGE
	AUTHOR	LANG_USAGE_REF
	TERM	LANGUSE_REF_CONF
	TERM_TYPE	
	LOOKUP_FORM	REGIONAL_USAGE
	OBSOLETE	REG_USAGE_REF
		REGUSE_REF_CONF
	TL_COMMENT	
	COMMENT_CONF	CONTEXT
	DATE_COMMENT_MADE_CONF	CONTEXT_REF
	COMMENT_MADE_CONF_BY_USER	CONTEXT_REF_CONF
		GENDER
	RELIABILITY_VALUE	PART_OF_SPEECH
	TERM_REF	
	TERM_REF_CONF	

Term level (includes word level information)	TL_RECORD AUTHOR PROPOSER INSTITUTION CREATION_DATE CHANGED_BY CHANGE_DATE	CHANGED_IN_FIELDS MARKED_FOR_DELETION_MERGING INITIAL_SOURCE VALIDATION_STATUS STAGE CYCLE

APPENDIX 2. EXAMPLES OF DIFFERENCES IN CONTENT AND STRUCTURE OF TERMINOLOGICAL RESOURCES FROM EUROTERMBANK SOURCE DATA



L 361 lähenema 106 L 388 lüliti

L 361	lähenema piirväärtusele	approach a limit	приближаться к пределу
L 362	lähenemine	approaching	приближение
L 363	lähim	closest, nearest	ближайший
L 364	lähis-	adjacent	прилежащий
L 365	lähis- (= lähend-)	approximate	приближённый
L 366	lähiskaatet	adjacent side	прилежащий катет
L 367	lähiskülg	adjacent side	прилежащая сторона
L 368	lähismurd	convergent	подходящая дробь
L 369	lähte-	initial, original, starting	исходный, начальный
L 370	lähteandmed (⇒ algandmed)	initial data, source data	исходные данные, начальные данные
L 371	lähtehulk	domain, initial set	множество отправления
L 372	lähtepunkt (= alguspunkt)	initial point, starting point	исходная точка, начальная точка, отправная точка
L 373	lähtesümbol	initial symbol	начальный символ
L 374	lähtevõrrand	original equation	исходное уравнение, первоначальное уравнение
L 375	lähtuma	start	исходить
L 376	lävi	threshold	порог
L 377	lühem telg (⇒ väike-telg)	minor axis	малая ось
L 378	lühend	abbreviation	аббревиатура, сокращение
L 379	lühendama	abbreviate, shorten	сокращать, укорачивать
L 380	lühendamine	abbreviation, shortening	сокращение, укорачивание
L 381	lühendatud	abbreviated, abridged, contracted	сокращённый
L 382	lühendatud korrutamine	abbreviated multiplication	сокращённое умножение
L 383	lühendatud tähis	abridged notation, contracted notation	сокращённое обозначение
L 384	lühike	short	короткий
L 385	lühim	shortest	кратчайший
L 386	lüke (⇐ translatsioon)	shift, slide, translation	перемещение, перенос, сдвиг
L 387	lüli	link	звено
L 388	lüliti	switch	переключатель

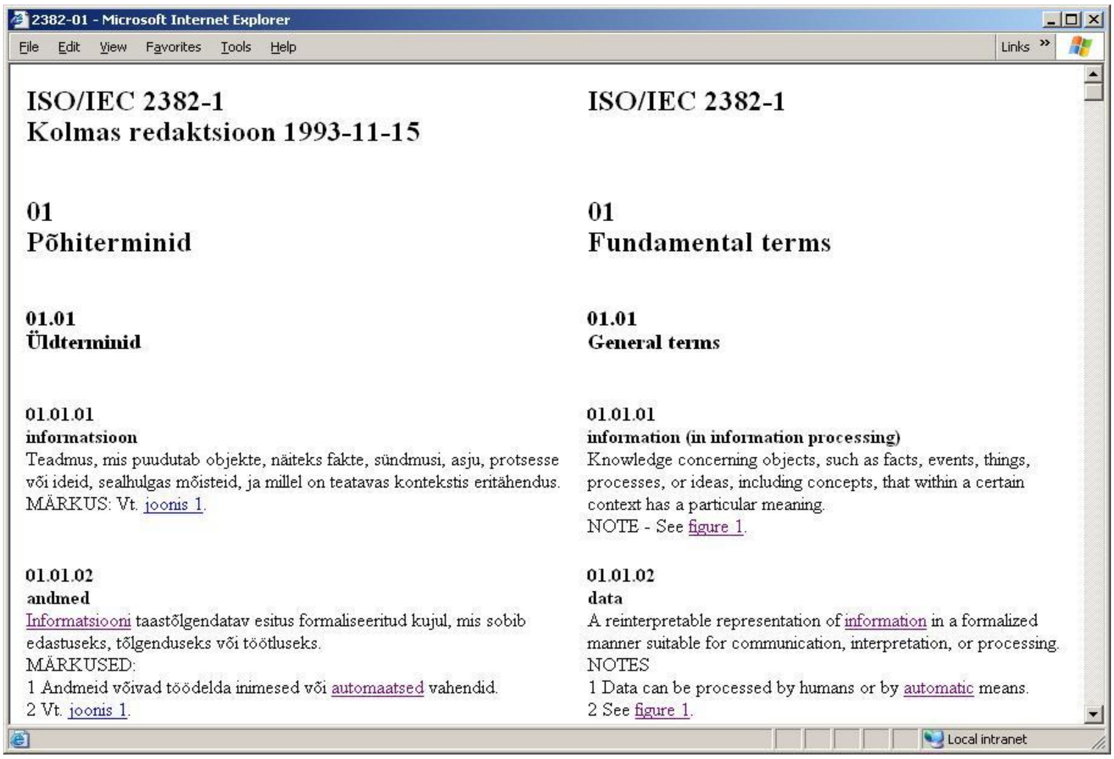
M 1 maagiline 107 M 33 majoreerimine

M

M 1	maagiline ruut	magic square	магический квадрат
M 2	maastikuformaat	landscape size	горизонтальный формат
M 3	maatriks	matrix	матрица
M 4	maatriks-	matrix	матричный
M 5	maatriksalgebra	matrix algebra	матричная алгебра
M 6	maatriksesitus	matrix representation	матричное представление
M 7	maatriksi astak	rank of a matrix	ранг матрицы
M 8	maatriksi diagonaal	diagonal of a matrix	диагональ матрицы
M 9	maatriksi elementaarteisendused	elementary operations on matrices	элементарные операции над матрицами
M 10	maatriksi jälg	spur of a matrix, trace of a matrix	след матрицы
M 11	maatriksi pööramine	inversion of a matrix, matrix inversion	обращение матрицы
M 12	maatriksite liitmine	addition of matrices	сложение матриц
M 13	maatriksite ring	ring of matrices	кольцо матриц
M 14	maatriksmäng	matrix game	матричная игра
M 15	maatrikstähistus	matrix notation	матричная символика
M 16	maatriksväärtustega funktsioon	matrix-valued function	матрица-функция, матричнозначная функция
M 17	Mackey topoloogia	Mackey topology	топология Макки
M 18	Maclaurini rida	Maclaurin's series	ряд Маклорена
M 19	madalaim numbrikoht	least significant digit	младший разряд
M 20	madalamat järku	of lower order	низшего порядка
M 21	magasin (⇐ pinu)	stack	магазин, стек
M 22	magasinalgoritm	stack algorithm	магазинный алгоритм
M 23	magasinautomaat	push-down automaton	магазинный автомат
M 24	magasini tipp	top of stack	вершина стека
M 25	magistraal	turnpike	магистраль
M 26	maha tõmbama	cross out	вычёркивать
M 27	mahatõmbamine	crossing out	вычёркивание
M 28	maht (= mahukus)	capacity	ёмкость
M 29	mahukas	capacious	ёмкий
M 30	mahukeskmine	power mean	степенное среднее
M 31	majorant	majorant	мажоранта
M 32	majoreerima	majorize	мажорировать
M 33	majoreerimine	majorizing	мажорирование

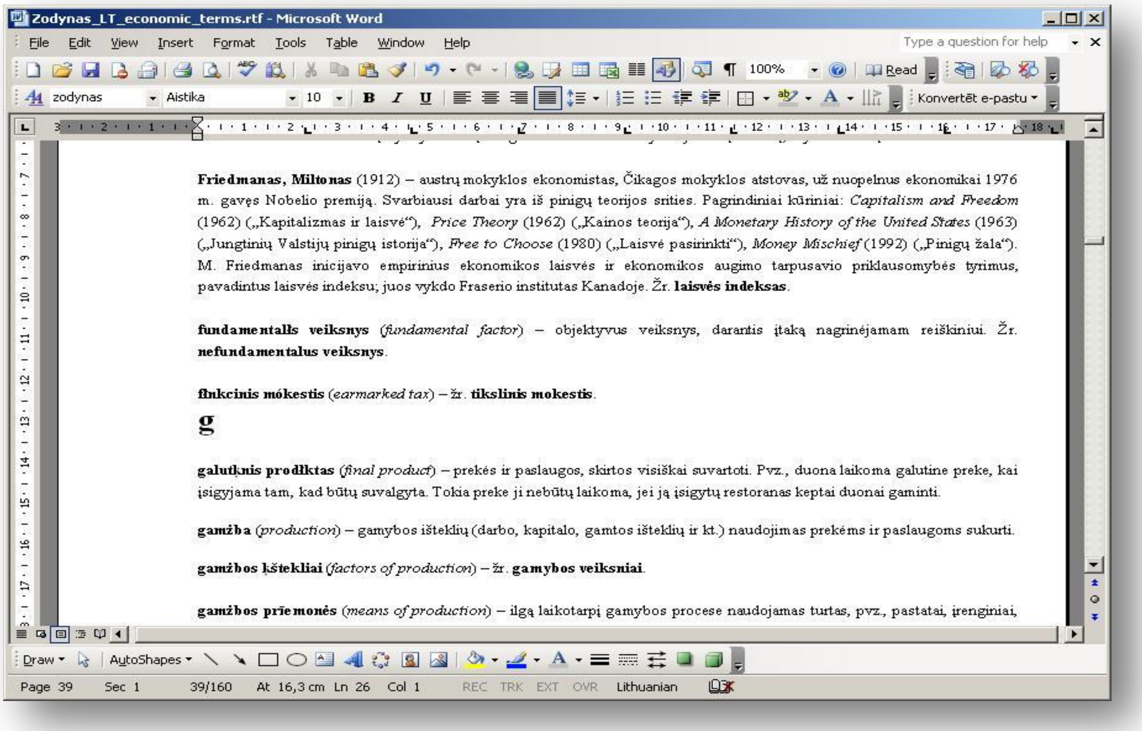
14*

EE-EN-RU Mathematical Term Collection



ISO/IEC 2382-1 Kolmas redaktsioon 1993-11-15	ISO/IEC 2382-1
01 Põhiterminid	01 Fundamental terms
01.01 Üldterminid	01.01 General terms
01.01.01 informatsioon Teadmus, mis puudutab objekte, näiteks fakte, sündmusi, asju, protsesse või ideid, sealhulgas mõisteid, ja millel on teatavas kontekstis eritähendus. MÄRKUS: Vt. joonis 1.	01.01.01 information (in information processing) Knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning. NOTE - See figure 1.
01.01.02 andmed Informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluseks. MÄRKUSED: 1 Andmeid võivad töödelda inimesed või automaatsed vahendid. 2 Vt. joonis 1.	01.01.02 data A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. NOTES 1 Data can be processed by humans or by automatic means. 2 See figure 1.

EN-LV Term Financial Term Collection


Zodynas_LT_economic_terms.rtf - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help
100%
Read
zodynas
Aistika
10
Konvertēt e-pastu
Draw
AutoShapes
Page 39 Sec 1 39/160 At 16,3 cm Ln 26 Col 1 REC TRK EXT OVR Lithuanian

Friedmanas, Miltonas (1912) – austrų mokyklos ekonomistas, Čikagos mokyklos atstovas, už nuopelnus ekonomikai 1976 m. gavęs Nobelio premiją. Svarbiausi darbai yra iš pinigų teorijos srities. Pagrindiniai kūriniai: *Capitalism and Freedom* (1962) („Kapitalizmas ir laisvė“), *Price Theory* (1962) („Kainos teorija“), *A Monetary History of the United States* (1963) („Jungtinių Valstijų pinigų istorija“), *Free to Choose* (1980) („Laisvė pasirinkti“), *Money Mischief* (1992) („Pinigų žala“). M. Friedmanas inicijavo empirinius ekonomikos laisvės ir ekonomikos augimo tarpusavio priklausomybės tyrimus, pavadintus laisvės indeksu; juos vykdo Fraserio institutas Kanadoje. Žr. **laisvės indeksas**.

fundamentalłs veiksnys (*fundamental factor*) – objektyvus veiksnys, darantis įtaką nagrinėjamam reiškiniui. Žr. **nefundamentalus veiksnys**.

fłnkcinis mókestis (*earmarked tax*) – žr. **tikslinis mokestis**.

g

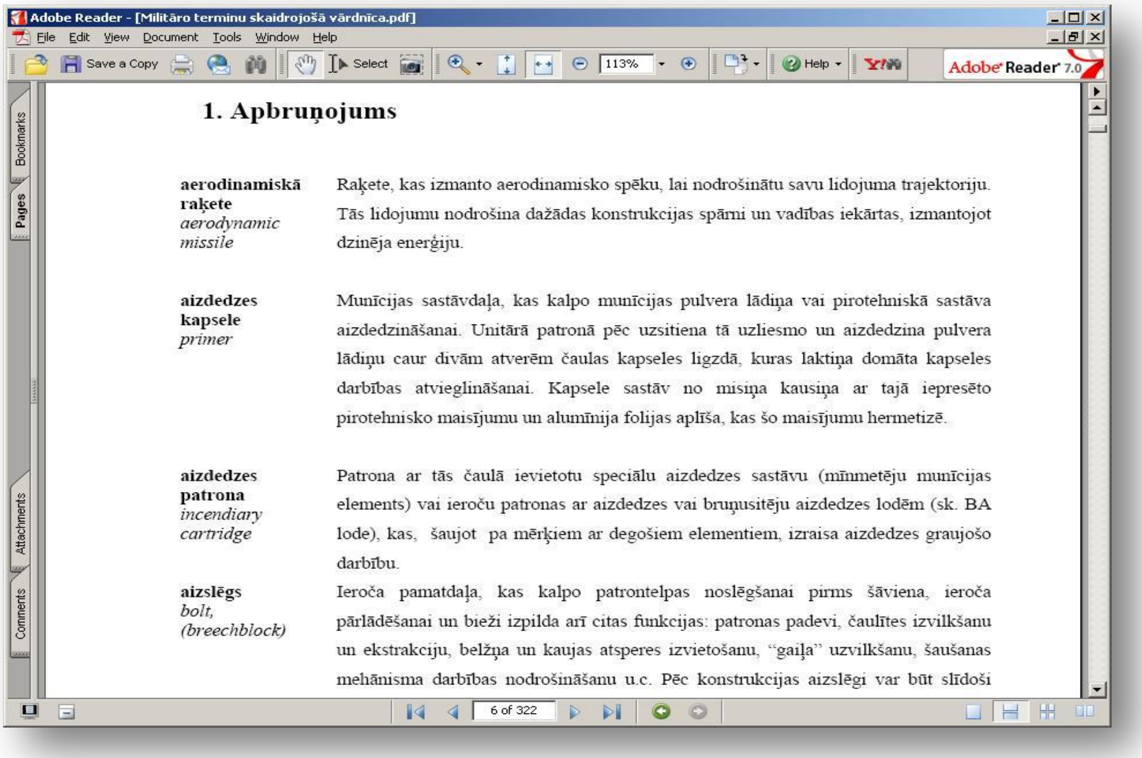
galutķnis prodłktas (*final product*) – prekės ir paslaugos, skirtos visiškai suvartoti. Pvz., duona laikoma galutine preke, kai įsigyjama tam, kad būtų suvalgyta. Tokia preke ji nebūtų laikoma, jei ją įsigytų restoranas keptai duonai gaminti.

gamżba (*production*) – gamybos išteklių (darbo, kapitalo, gamtos išteklių ir kt.) naudojimas prekėms ir paslaugoms sukurti.

gamżbos ķštekliai (*factors of production*) – žr. **gamybos veiksniai**.

gamżbos prīemonės (*means of production*) – ilgą laikotarpį gamybos procese naudojamas turtas, pvz., pastatai, įrenginiai,

LT Explanatory Dictionary of Economical Terms


Adobe Reader - [Militāro terminu skaidrojošā vārdnīca.pdf]
File Edit View Document Tools Window Help
Save a Copy
Select
113%
Help
Adobe Reader 7.0
Bookmarks
Pages
Attachments
Comments
6 of 322

1. Apbruņojums

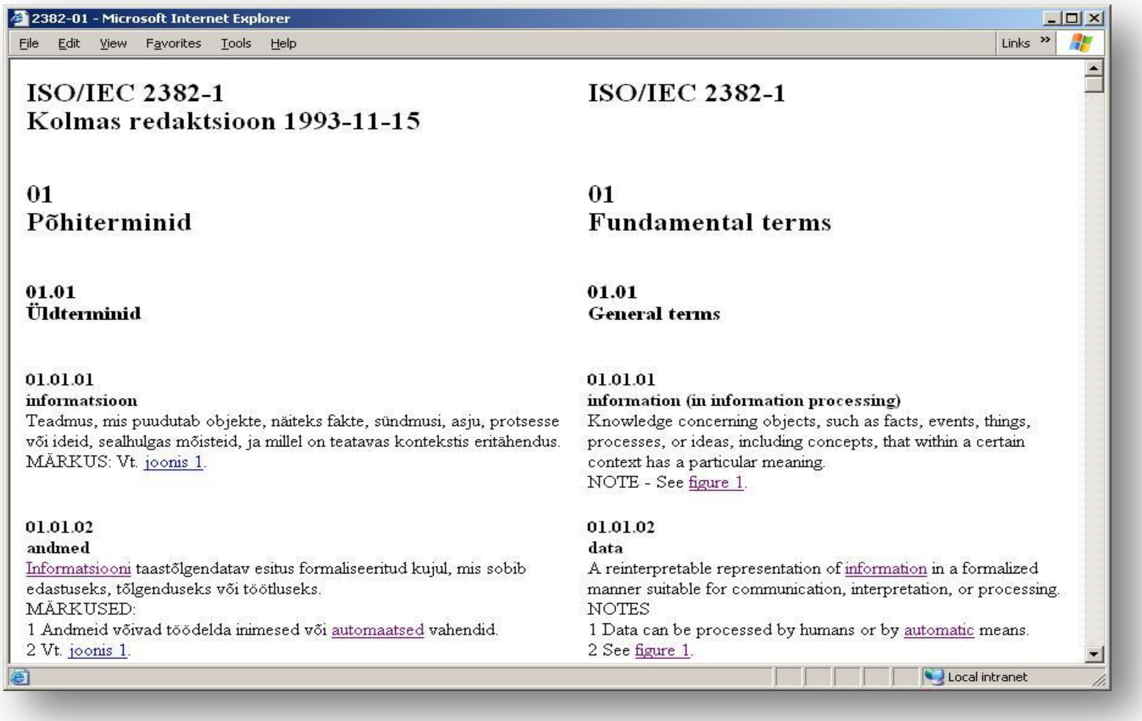
aerodinamiskā raķete *aerodynamic missile*
Raķete, kas izmanto aerodinamisko spēku, lai nodrošinātu savu lidojuma trajektoriju. Tās lidojumu nodrošina dažādas konstrukcijas spārni un vadības iekārtas, izmantojot dzinēja enerģiju.

aizdedzes kapsele *primer*
Munīcijas sastāvdaļa, kas kalpo munīcijas pulvera lādiņa vai pirotehniskā sastāva aizdedzināšanai. Unitārā patronā pēc uzsitiena tā uzliesmo un aizdedzina pulvera lādiņu caur divām atverēm čaulas kapseles ligzdā, kuras laktiņa domāta kapseles darbības atvieglināšanai. Kapsele sastāv no misiņa kausiņa ar tajā iepresēto pirotehnisko maisījumu un alumīnija folijas aplīša, kas šo maisījumu hermetizē.

aizdedzes patrona *incendiary cartridge*
Patrona ar tās čaulā ievietotu speciālu aizdedzes sastāvu (mīnmetēju munīcijas elements) vai ieroču patronas ar aizdedzes vai bruņusitēju aizdedzes lodēm (sk. BA lode), kas, šaujot pa mērķiem ar degošiem elementiem, izraisa aizdedzes graujošo darbību.

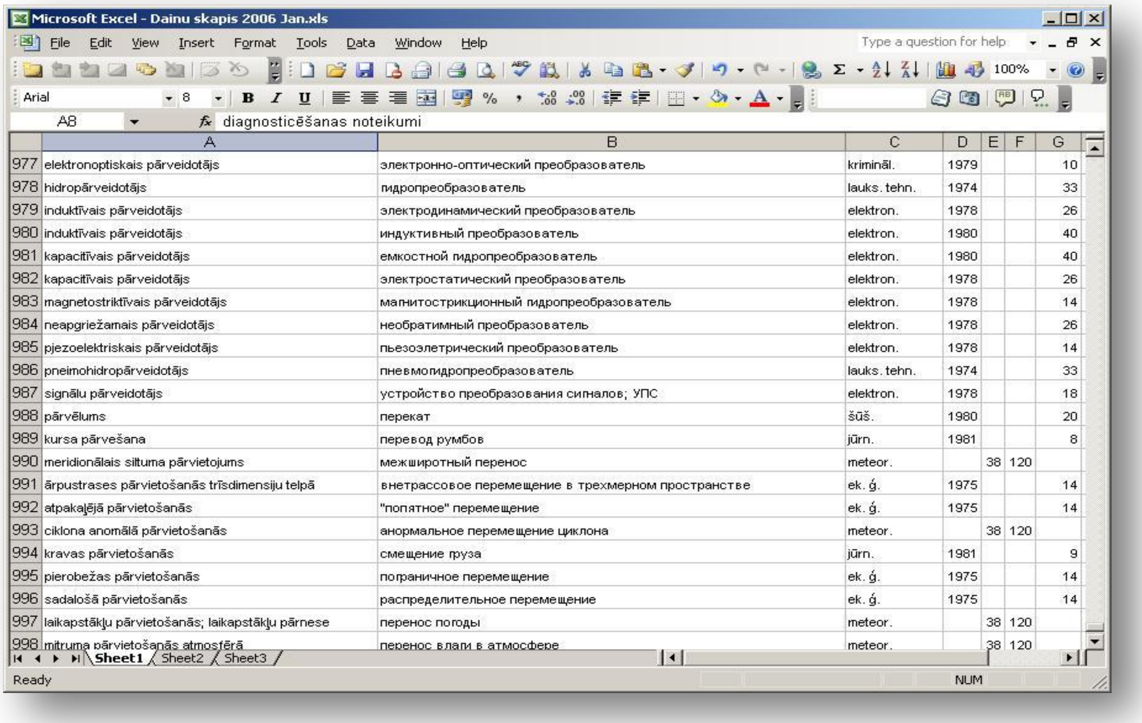
aizslēgs *bolt, (breechblock)*
Ieroča pamatdaļa, kas kalpo patrontelpas noslēgšanai pirms šāviena, ieroča pārlādēšanai un bieži izpilda arī citas funkcijas: patronas padevi, čaulītes izvilkšanu un ekstrakciju, belžņa un kaujas atsperes izvietošanu, "gaiļa" uzvilkšanu, šaušanas mehānisma darbības nodrošināšanu u.c. Pēc konstrukcijas aizslēgi var būt slīdoši

LV Military Term Dictionary


2382-01 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Local intranet

ISO/IEC 2382-1 **Kolmas redaktsioon 1993-11-15**	**ISO/IEC 2382-1**
01 **Põhiterminid**	**01** **Fundamental terms**
01.01 **Üldterminid**	**01.01** **General terms**
01.01.01 **informatsioon** Teadmus, mis puudutab objekte, näiteks fakte, sündmusi, asju, protsesse või ideid, sealhulgas mõisteid, ja millel on teatavas kontekstis eritähendus. MÄRKUS: Vt. joonis 1.	**01.01.01** **information (in information processing)** Knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning. NOTE - See figure 1.
01.01.02 **andmed** Informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluseks. MÄRKUSED: 1 Andmeid võivad töödelda inimesed või automaatsed vahendid. 2 Vt. joonis 1.	**01.01.02** **data** A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. NOTES 1 Data can be processed by humans or by automatic means. 2 See figure 1.

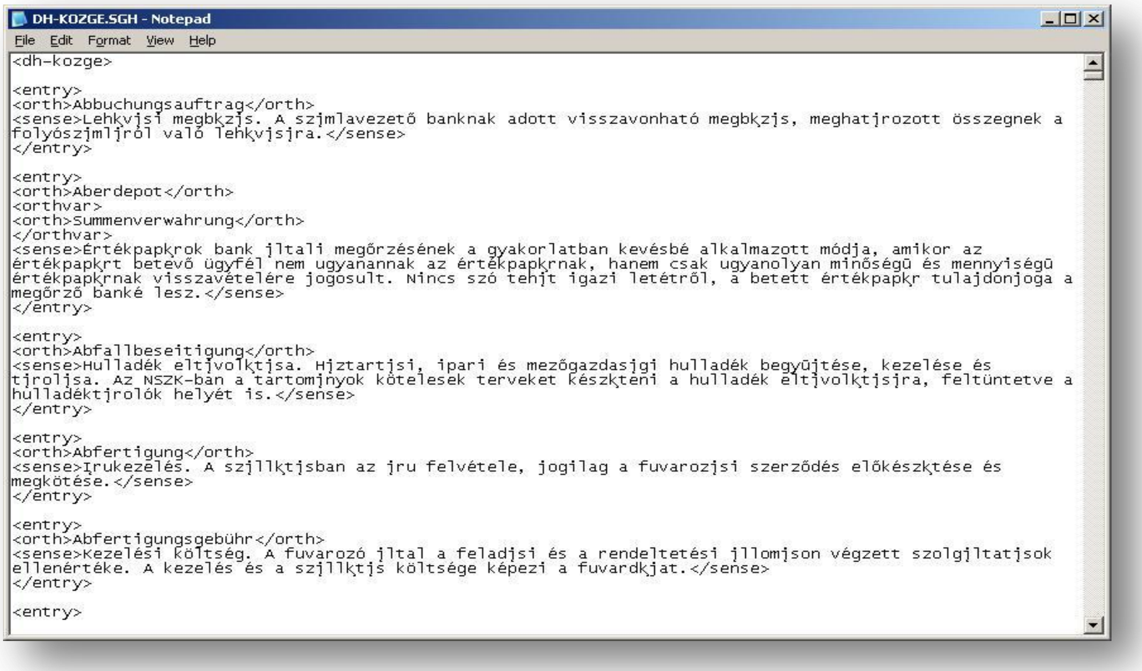
EE ISO IT Standard Terminology


Microsoft Excel - Dainu skapis 2006 Jan.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
100%
Arial
8
Sheet1
Sheet2
Sheet3
Ready
NUM

A8 diagnosticēšanas noteikumi

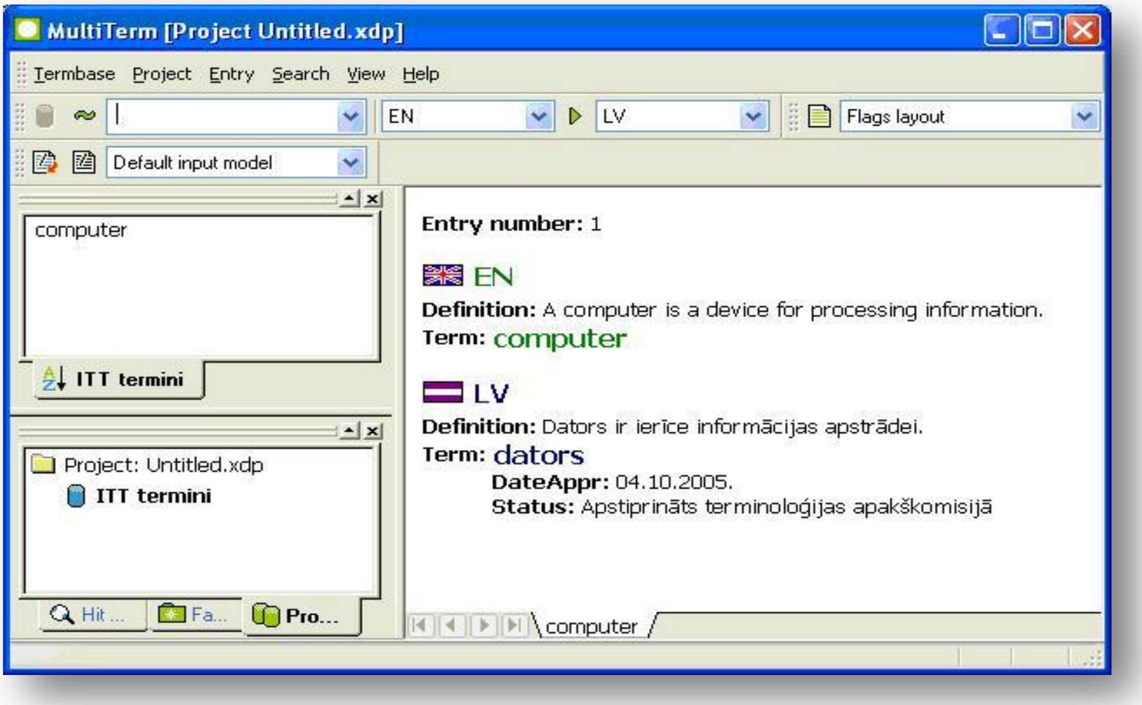
	A	B	C	D	E	F	G
977	elektronoptiskais pārveidotājs	электронно-оптический преобразователь	kriminā l.	1979			10
978	hidropārveidotājs	гидропреобразователь	lauks. tehn.	1974			33
979	induktīvais pārveidotājs	электродинамический преобразователь	elektron.	1978			26
980	induktīvais pārveidotājs	индуктивный преобразователь	elektron.	1980			40
981	kapacitīvais pārveidotājs	емкостной гидропреобразователь	elektron.	1980			40
982	kapacitīvais pārveidotājs	электростатический преобразователь	elektron.	1978			26
983	magnetostriktīvais pārveidotājs	магнитострикционный гидропреобразователь	elektron.	1978			14
984	neapgriežamais pārveidotājs	необратимный преобразователь	elektron.	1978			26
985	pjezoelektriskais pārveidotājs	пьезоэлетрический преобразователь	elektron.	1978			14
986	pneimohidropārveidotājs	пневмогидропреобразователь	lauks. tehn.	1974			33
987	signālu pārveidotājs	устройство преобразования сигналов; УПС	elektron.	1978			18
988	pārvēlums	перекат	šūš.	1980			20
989	kursa pārvešana	перевод румбов	jūrn.	1981			8
990	meridionālais siltuma pārvietojums	межширотный перенос	meteor.		38	120	
991	ārpustrases pārvietošanās trīsdimensiju telpā	внетрассовое перемещение в трехмерном пространстве	ek. ģ.	1975			14
992	atpakaļējā pārvietošanās	"попятное" перемещение	ek. ģ.	1975			14
993	ciklona anomālā pārvietošanās	анормальное перемещение циклона	meteor.		38	120	
994	kravas pārvietošanās	смещение груза	jūrn.	1981			9
995	pierobežas pārvietošanās	пограничное перемещение	ek. ģ.	1975			14
996	sadalošā pārvietošanās	распределительное перемещение	ek. ģ.	1975			14
997	laikapstākļu pārvietošanās; laikapstākļu pārnese	перенос погоды	meteor.		38	120	
998	mitruma pārvietošanās atmosfērā	перенос влаги в атмосфере	meteor.		38	120	

LV-RU Cross-disciplinary Terminology



```
<dh-kozge>

<entry>
<orth>Abbuchungsauftrag</orth>
<sense>Lehķvjsi megbķzjs. A szjmlavezető banknak adott visszavonható megbķzjs, meghatjrozott összegnek a
folyószjmljról való lehķvjsjra.</sense>
</entry>

<entry>
<orth>Aberdepot</orth>
<orthvar>
<orth>Summenverwahrung</orth>
</orthvar>
<sense>Értékpapķrok bank jltali megőrzésének a gyakorlatban kevésbé alkalmazott módja, amikor az
értékpapķrt betevő ügyfél nem ugyanannak az értékpapķrnak, hanem csak ugyanolyan minőségū és mennyiségū
értékpapķrnak visszavételére jogosult. Nincs szó tehjt igazi letétről, a betett értékpapķr tulajdonjoga a
megőrző banké lesz.</sense>
</entry>

<entry>
<orth>Abfallbeseitigung</orth>
<sense>Hulladék eltjvolķtjsa. Hjztartjsi, ipari és mezőgazdasjgi hulladék begyūjtése, kezelése és
tjroljsa. Az NSZK-ban a tartomjnyok kötelesek terveket készķteni a hulladék eltjvolķtjsjra, feltüntetve a
hulladéktjrolók helyét is.</sense>
</entry>

<entry>
<orth>Abfertigung</orth>
<sense>Įrukezelés. A szjllķtjsban az jru felvétele, jogilag a fuvarozjsi szerződés előkészķtése és
megkötése.</sense>
</entry>

<entry>
<orth>Abfertigungsgebühr</orth>
<sense>Kezelési költség. A fuvarozó jltal a feladjsi és a rendeltetési jllomjson végzett szolgjltatjsok
ellenértéke. A kezelés és a szjllķtjs költsége képezi a fuvardķjat.</sense>
</entry>

<entry>
```

DE-HU Financial Terminology



Entry number: 1

EN

Definition: A computer is a device for processing information.

Term: computer

LV

Definition: Dators ir ierīce informācijas apstrādei.

Term: dators

DateAppr: 04.10.2005.

Status: Apstiprināts terminoloģijas apakškomisijā

Possible representation of term entry in Trados MultiTerm environment

APPENDIX 3. THE ETB DATA EXCHANGE FORMAT

Tag & sample	ISO 12620	Required	Description
<?xml version=1.0 encoding=utf-8?>			
<!DOCTYPE martif PUBLIC ISO 12200:1999A//DTD MARTIF core (DXFcdV04)//EN TBXcdv04.dtd>			
<martif type=TBX xml:lang=en>			
<martifHeader>			
<fileDesc>			
<titleStmt>			
<title>Title of the collection</title>		*	Title of the collection
</titleStmt>			
<sourceDesc>			
<p>Description of the collection source</p>			Description of the source
</sourceDesc>			
</fileDesc>			
<encodingDesc>			
<p type=DCSName>TBXDv04C			File with encoding description

ycom.xml</p>			
</encodingDesc>			
</martifHeader>			
<text>			
<body>			
<termEntry id='ID67'>	A.10.15	*	Entry identifier a system-generated number that will identify the entry uniquely
<admin type='sourceLanguage'>en</admin	A.10.23		Source Language the source language of a set of terms that are not perfectly multi-directional
<admin type='subsetOwner'>SIA TILDE</admin>	A.10.02.02.10	*	Subset owner institution responsible for the whole entry
<admin type='securitySubset'>2</admin>	A.10.03.09	*	Security subset a security classification expressing the confidentiality level of the entire entry
<transacGrp>		*	
<transac type=transactionType>origination</transac>		*	
<transacNote type='responsibility'>R. Smith</transacNote>	A.10.02.02.01	*	Originator an identifier of the person who prepared the entry
<date></date>	A.10.02.01.0	*	Origination date

	1		The date the entry was first created
</transacGrp>		*	
<transacGrp>		*	
<transac type=transactionType>creation </transac>		*	
<transacNote type='responsibility'>J. Smith</transacNote>	A.10.02.02.02	*	Inputter An identifier of the person who types in the information
</transacGrp>		*	
<transacGrp>			
<transac type=transactionType>modification</transac>			
<transacNote type='responsibility'>J. Clarck</transacNote>	A.10.02.02.03		Updater the person having made the latest changes to the information at entry level
<date></date>	A.10.02.01.03		Modification date The date when the latest changes to the entry level were made
</transacGrp>			
<descrip type='subjectField'>23</descrip>	A.04	*	Subject Field the subject of the concept
<note>more subject	A.08		Note

information</note>			a note related to the *classification number*
<descrip type='otherBynaryData'>235j239sd21</descrip>	A.05.05.05		Other binary data
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference
<admin type='projectSubset'>abc</admin>	A.10.03.03		Project subset an identifier of a particular collection of concepts
<descrip type='broaderConceptGeneric' target='entryId'> </descrip>	A.07.02.01		Broader concept
<descrip type='subordinateConceptGeneric' target='entryId'></descrip>	A.07.02.03		Subordinate concept
<descrip type='relatedConcept' target='entryId'></descrip>	A.07.02.05		Related concept
<langSet lang=en'>	A.10.07	*	Language symbol the language symbol of the particular language
<transacGrp>		*	
<transac type=transactionType>origination</transac>		*	

<transacNote type='responsibility'>R. Smith</transacNote>	A.10.02.02.01	*	Originator an identifier of the person who prepared the language level
<date></date>	A.10.02.01.01	*	Origination date The date the language level was first created
</transacGrp>		*	
<transacGrp>		*	
<transac type=transactionType>creation </transac>		*	
<transacNote type='responsibility'>J. Smith</transacNote>	A.10.02.02.02	*	Inputter An identifier of the person who types in the information
</transacGrp>		*	
<transacGrp>			
<transac type=transactionType>modification</transac>			
<transacNote type='responsibility'>J. Clarck</transacNote>	A.10.02.02.03		Updater the person having made the latest changes to the information at language level
<date></date>	A.10.02.01.03		Modification date The date when the latest changes to the language level were made
</transacGrp>			

<descrip type='otherBynaryData'>235j239sd21</descrip>	A.05.05.05		Other binary data
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference
<note>more inf about the concept in particular language</note>	A.08		Note A note field related to the entire language level
<descrip type='reliabilityCode'>2</descrip>	A.03.04		Reliability code an assessment of the correctness and precision of the information given in relation to the specific concept.
<descripGrp>			
<descrip type='definition'>degree of obstruction</descrip>	A.05.01		Definition
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference a reference to the definition
</descripGrp>			
<descripGrp>			
<descrip type='explanation'>degree of obstruction</descrip>	A.05.02		Explanation
<admin	A.10.20		Reference

type='sourceIdentifier' target='DIN-561.12'>p.21</ref>			A reference to the explanation
</descripGrp>			
<ntig>			
<transacGrp>			
<transac type=transactionType>origination</transac>			
<transacNote type='responsibility'>R. Smith</transacNote>	A.10.02.02.01	*	**Originator -** an identifier of the person who prepared the term level
<date></date>	A.10.02.01.01	*	**Origination date -** The date the term level was first created
</transacGrp>			
<transacGrp>			
<transac type=transactionType>creation</transac>			
<transacNote type='responsibility'>J. Smith</transacNote>	A.10.02.02.02	*	**Inputter -** An identifier of the person who types in the information
</transacGrp>			
<transacGrp>			
<transac type=transactionType>modification</transac>			
<transacNote	A.10.02.02.0		**Updater** – the person having

type='responsibility'>J. Clarck</transacNote>	3		made the latest changes to the information at term level
<date></date>	A.10.02.01.03		**Modification date** - The date when the latest changes to the term level were made
</transacGrp>			
<transacGrp>			
<transac type=transactionType>approval</transac>			
<transacNote type='responsibility'>R. Smith</transacNote>	A.10.02.02.04		**Approver** – An identifier of the person consolidating the entry
<date></date>	A.10.02.01.04		Approval date
</transacGrp>			
<termGrp>			
<admin type='entrySource'>db</admin>	A.10.13		Entry source the database or format from which data are imported
<admin type='intellectualPropertyRights'>p.21</admin>	No ISO Code		Intellectual property rights
<descrip type='context'>state transition table</descrip>	A.05.03		Context
<admin type='sourceIdentifier' target='DIN-	A.10.20		Reference Source(s) of the context

561.12'>p.21</ref>			example
<termNote type='register' >neutralRegister</termNote>	A.02.03.03		Register a classification indicating the relative level of language assigned to a term
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference Reference(s) to the *register* information
<termNote type='temporalQualifier' >archaicTerm</termNote>	A.02.03.05		Temporal qualifier Information about a term with respect to its use over time
<termNote type='usageNote' >rarely used</termNote>	A.02.03.01		Usage note local, regional or geographic usage of the term
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference Reference(s) to the *Usage note* field.
<note>general note to term level</note>	A.08		Note A general comment that applies to the entire term level
<descrip type='reliabilityCode'>4</descrip>	A.03.04		Reliability code an assessment of the correctness and precision of the information given in relation to the specific term
<termNote	A.02.09.01		Normative authorization

type='normativeAuthorization' >preferredTerm</termNote>			A term status qualifier assigned by an authoritative body
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20		Reference Reference to the normative organization
<admin type='searchTerm'>transition table</admin>	A.10.06.03		Search term related forms of the term to facilitate searching
<term>transition table</term>	A.01	*	Term
<termNote type='termType' >fullForm</termNote>	A.02.01	*	Term Type Some possible values are: main entry term, abbreviation, acronym, short form, variant, formula, synonym ….
<admin type='sourceIdentifier' target='DIN-561.12'>p.21</ref>	A.10.20	*	Reference Source(s) of the *term*.
<termCompList type=termElement>			
<transacGrp>		*	
<transac type=transactionType>origination</transac>		*	
<transacNote type='responsibility'>R.	A.10.02.02.01	*	Originator an identifier of the person who

Smith</transacNote>			prepared the word level
<date></date>	A.10.02.01.01	*	Origination date The date the word level was first created
</transacGrp>		*	
<transacGrp>		*	
<transac type=transactionType>creation </transac>		*	
<transacNote type='responsibility'>Smith</transacNote>	A.10.02.02.02	*	Inputter An identifier of the person who types in the information
<trnsacGrp>		*	
<transacGrp>			
<transac type=transactionType>modification</transac>			
<transacNote type='responsibility'>Clarck</transacNote>	A.10.02.02.03		Updater the person having made the latest changes to the information at word level
<date></date>	A.10.02.01.03		Modification date The date when the latest changes to the word level were made
</transacGrp>			
<termCompGrp>			

<termComp>transition</termComp>	A.02.08.02		Term element a particular word that forms part of a term
<termNote type=partOfSpeech>noun</termNote>	A.02.02.01		Part of speech
<termNote type=grammaticalNumber>singular</termNote>	A.02.02.03		Grammatical number
<termNote type=grammaticalGender>masculine</termNote>	A.02.02.02		Grammatical gender
<termCompList type=morphologicalElement>some other morph info</termCompList>	A.02.08.01		Morphological element
<termNote type=pronunciation>…</termNote>	A.02.05		Pronunciation Pronunciation information like accentuation of syllables
</termCompGrp>			
...	…		Other word level items follow here
</termCompList>			
</termGrp>			
</ntig >			
...	…		Other terms follow here
</langSet>			

…	…		Other language level records follow here
</termEntry>			
</body>			
<back>			
<refObjectList type=bibl>			Description of the references used in file
<refObject id=piggott97>			Reference object with its identifier
<itemSet type=article>			Type of the reference
<item type=title>Glossary</item>			Title of the reference
</itemSet>			
<itemSet type=author>			
<item type=surname>Piggott</item>			Last name of the author
<item type=fname>Hugh</item>			First name of the author
</itemSet>			
<itemSet type=book>			Type of the reference source
<item type=title>Windpower workshop</item>			Title of the source
<item type=edition>First</item>			Edition of the source
<item type=isbn>1 898049 13 0</item>			ISBN of the source

</itemSet>			
<item type=pages>138-144</item>			Pages of the source
<item type=date>1997-05</item>			Date of the source
<itemSet type=pubname>			Publisher information
<item type=orgName>The Centre for Alternative Technology</item>			Publisher organization name
</itemSet>			
</refObject>			
...			Other reference objects follow here
</refObjectList>			
</back>			
</text>			
</martif>			