META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D3.6

Multilingual Terminology

Version No. 1.0
31/01/2013

Document Information

Deliverable number:	D3.6
Deliverable title:	Multilingual Terminology
Due date of deliverable:	31/01/2013
Actual submission date of deliverable:	31/01/2013
Main Author(s):	Tilde
Participants:	All partners
Internal reviewer:	UGOT
Workpackage:	WP3
Workpackage title:	Enhancing language resources
Workpackage leader:	Tilde
Dissemination Level:	PU
Version:	1.0
Keywords:	Terminology, resources, metadata

EXECUTIVE SUMMARY

This report documents the activities performed in T3.6 – Horizontal Action on multilingual terminology. In separate sections of this document we describe the activities with terminology resources – identifying, collecting, and licensing for META-SHARE. For building a common linguistic infrastructure, the EuroTermBank terminology database was interlinked with META-SHARE, allowing resource metadata harvesting from language resource specific nodes for the META-SHARE network. This resulted in 99 new resources added to META-SHARE with a live link to the source databank. The last section covers possible and real synergies scenarios based on the language resource specific database integration with META-SHARE, sought during the META-NORD project which are very likely to be implemented shortly after the project.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LR	Language Resource
MT	Machine Translation
TBX	Term Base eXchange – ISO 30042 / Localization Industry Standard
INESS	The *Norwegian Infrastructure for the Exploration of Syntax and Semantics –* an eScience laboratory for linguistic research based on treebanks.
UGOT	University of Gothenburg

Introduction

Multilingual terminology consists of multiple terminology collections and databases of terms and concepts that in specific contexts and domains have specific meanings which may be different from the meaning in everyday life. The rapid development of technologies leads to the forming of new terms and concepts to describe them. Availability of up-to-date terminology allows professionals to communicate effectively with other professionals and non-professionals. Availability of terminology is crucial for translators, professionals, writers of technical documents, language students, MT systems etc. Unfortunately, terminology work is performed differently from country to country; with a few exceptions it is decentralized, often hidden, undisclosed, made confidential by industry key players, or access to it is hindered by a price tag.

This task addressed a growing demand to consolidate distributed terminology resources across languages and domains. META-NORD has extended an open linguistic infrastructure with multilingual terminology resources. The terminology consolidation platform EuroTermBank formerly established by the META-NORD partners Tilde, LKI, UT, and UCPH has been extended from 2 to 3 million terms internally by adding 43 terminology collections to EuroTermBank – LKI (Lithuanian) terminology, EASTIN terminology of Assistive Technology, and 40 collections from the Icelandic Terminology Bank. Three other terminology collections remain in the negotiation stage. One external term base (BFT, Finnish term bank) was connected to EuroTermBank during the META-NORD project, two other (STRUNA, Croatian term bank, PIRARC – Multilingual database of Road Terms) remain in the negotiation stage. Finally, a working prototype of interconnection of the EuroTermBank terminology resources has been created to link them with META-SHARE, allowing a content-specific Language Resource repository to be interconnected live with META-SHARE. During the META-NORD project this concept was replicated already – INESS resources of Treebanks were connected to META-SHARE via this proxy, too.

1. Interlinked Termbases

Before and during the project, the partners of META-NORD made their best effort to assess any available terminology resources in each country. Professional contacts and internet as a source of links and information was put to use. The base for the project's activity was the resource list which was part of the project Description of Work.

In co-location with TKE 2012, Tilde organized a public workshop on Creation, Harmonization and Application of Terminology Resources, CHAT 2012 [1] where terminology work of META-NORD and META-SHARE was presented. The workshop was targeted at holders of terminology resources, and some representatives of such institutions participated in it, too. The workshop helped us to get the word out on terminology work of META-NORD to a very broad international audience. Furthermore, Tilde was able to propose clarifications to the updated version of the TBX specification which was also presented and discussed at CHAT 2012, which is essential for terminology interchange between systems.

During the project, contacting and negotiations were made with resource owners with very different results. Although the goals of META-NORD and META-SHARE were received with understanding and generally positive attitudes, there were resource holders that did and some that did not get involved. The activity per country is as follows:

Iceland: The most spectacular, outstanding result was achieved by Icelandic partners who succeeded at negotiating with 40 authors whose work is contained in the **Icelandic Terminology Bank**. The rights to share for download were negotiated through a long and exhaustive process of negotiations. The resulting content was converted to the industry standard TBX format of terminology interchange and provided for download via META-SHARE network. In addition, the Icelandic Terminology Bank has been imported in EuroTermBank for centralized online lookup by users of the EuroTermBank terminology portal.

Norway: The UiB team has attempted to recover several terminological resources. The term bank Snorre has been reported in META-SHARE, and has been made available via the Norwegian National Library's digital language resource collection, Språkbanken. However, as regards terminology for the economical domain, the identified resources had a) been not been updated for a considerable period and would therefore require a considerable effort to be delivered in time, or b) were not mature enough to be included by the end of the project. This latter resource is owned by the Norwegian School of Economics (NHH) who is a partner in CLARINO with UiB. It was therefore decided to concentrate efforts on terminology in CLARINO, and prioritize META-NORD work on resources that would not be made available in other projects.

Sweden: Several attempts were made to locate and contact terminology holders in Sweden. TNC members participated in the CHAT seminar and seemed positive about granting their Swedish terms in some way to META-SHARE. However, it did not turn into any real activity, and numerous further attempts to contact them were left without any response.

Denmark: Only recently, a project on establishing a national Danish term bank has been launched at DanTermCenteret, Copenhagen Business School. This project will coordinate efforts and ensure development and quality of Danish terminology in the future. Before this initiative, there were only limited, publicly available termbanks. There has not been a tradition of making terminology that belonged to private companies and institutions publicly available at a larger scale.

Finland: The following three terminology resources were identified by University of Helsinki: Bank of Finnish Terminology (BFT) in Arts and Sciences http://tieteentermipankki.fi/; Valter Government Termbank http://mot.kielikone.fi/mot/valter/netmot.exe?UI=envl and TEPA Termbank http://www.tsk.fi/tepa/netmot.exe?UI=engr maintained by Finnish Terminology Centre TSK. All of these resources have their own web UI and they are available for browsing and search from their own websites. BFT has been integrated for one-stop terminology search with EuroTermBank, and negotiations are on-going with Valter and TEPA.

Estonia: there is a known terminology resource, ESTERM which is hosted online at http://mt.legaltext.ee/esterm/. It is possible to get access to different versions of this resource via different queries, for example www.keeleveeb.ee has a version from 2005 in its united query.

As much as project partners in UT know, it is not possible to settle copyright to this resource today because it consists of parts, which have different owners and even worse, the owner of the older part is reorganised without having any licencing conditions.

Latvia: During the project, a new (not known until now) resource was discovered, Technical Dictionary of Road Terms. This terminology resource is available for online search at http://www.lvceli.lv/LV/vardnica/PIARC/, as a printed book and as an electronic terminology database. However, agreement was not reached about providing this resource for public download.

Another similar resource is PIARC multilingual Terminology database which consists of the **Technical Dictionary of Road Terms**, the **Lexicon of Road and Traffic Engineering**, and a number of **specialized dictionaries** in the field of roads and road transport. These terms are available for online search from PIARC website at http://www.piarc.org/en/Terminology-Dictionaries-Road-Transport-Roads/term-search/

Negotiations were started and general acceptance was achieved about the Croatian Database of Special Field Terminology STRUNA, http://struna.ihjj.hr/en/, for it to be interlinked with EuroTermBank as part of the META-NORD project. However, the practical part of that task was delayed, but will most likely be completed shortly after the end of the META-NORD project.

Lithuania: LKI provided its terminology database of terms in Lithuanian language, and they will be provided for download in META-SHARE as well as included in EuroTermBank for consolidated online search. Terminological Synonyms of the Lithuanian Language database is a product of LKI that comprises sets of Lithuanian synonyms has been expanded by adding English language counterparts to Lithuanian synonym sets. The database has also been updated and extended by including new material. Presence of bilingual data has enabled

sharing the resource within EuroTermBank framework, by uploading its data in bilingual TBX format per ETB specifications. The resource is available for download through the Meta Share system, as well as accessible on the web through a dedicated interface.

2. Integration of EuroTermBank with META-SHARE

META-SHARE harvests the inventories of LR repositories into master META-SHARE inventories. In essence, LR repositories are META-SHARE nodes hosted by the LR providers. LR owners and META-SHARE node holders can enter the LR metadata in the META-SHARE editor, thus it becomes available to the entire META-SHARE network.
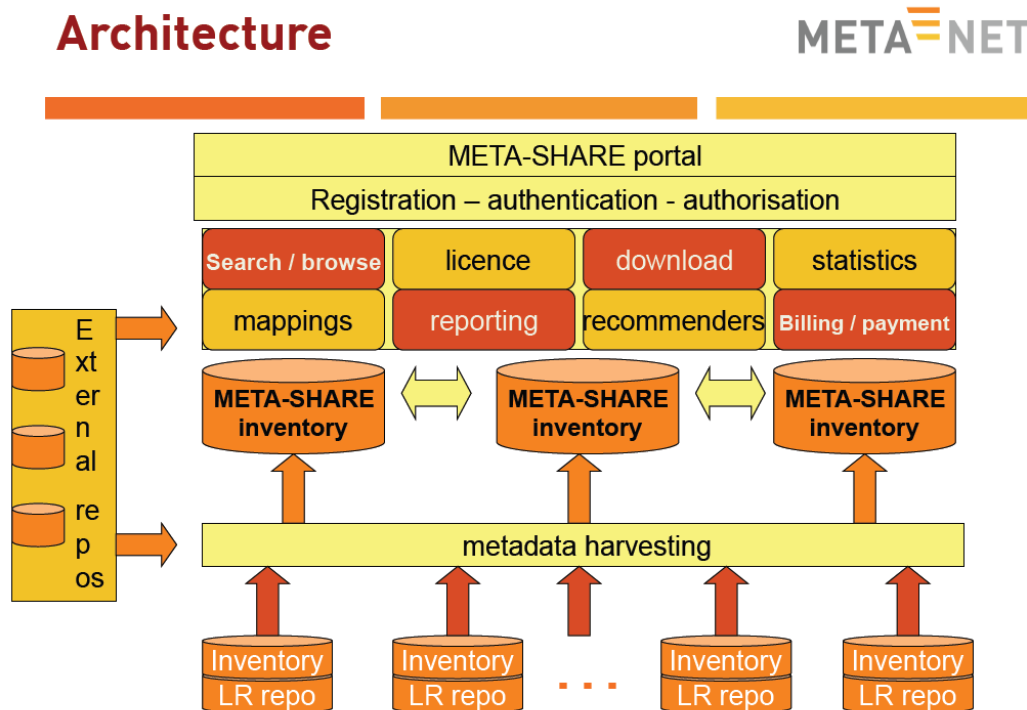

Figure 1 META-SHARE architecture

In real life there exist several LR-specific portals, such as term banks and tree-banks, which contain collections of LR resources where each individual resource is a candidate to be reflected in META-SHARE. It can be achieved by someone manually entering all the resource descriptions into the META-SHARE editor. The same can also be achieved via exporting the data from the respective portal, then converting it to META-SHARE schema compliant XML-s and importing it into any META-SHARE node. These approaches do not represent state-of-the-art solutions. They are time consuming; need regular updates; manual operations are error-prone, so they are not viable.

META-SHARE architecture is structured and open. In the META-NORD project, Tilde undertook to integrate EuroTermBank platform into an open linguistic infrastructure by adapting it to relevant data access and sharing specifications. This means that a language resource specific repository like EuroTermBank for terminology or any other could seamlessly become part of the META-SHARE network by adapting its data to the communication protocols, in the case of META-SHARE – to the communication of metadata harvesting. The diagram below shows how the inventory of the Terminology Repository is becoming part of the META-SHARE architecture.
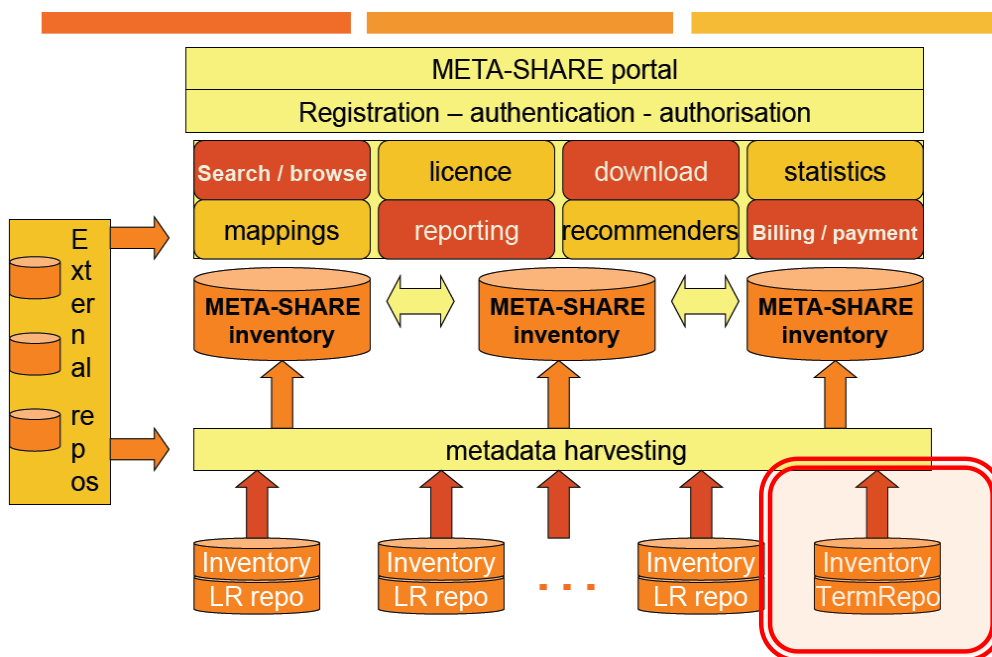
Architecture

Figure 2 Integration of LR specific repository in the META-SHARE architecture

Tilde approached the development team of META-SHARE for assistance in implementing this concept. After several rounds of explaining the concept, META-SHARE shared documents with Tilde, META-NET D7.1.2 (Functional-Technical Specs) and META-NET D6.2.2 (Design-Governance), as well as a draft of the META-SHARE Harvesting Protocol v1.0. Tilde based its development on those documents.
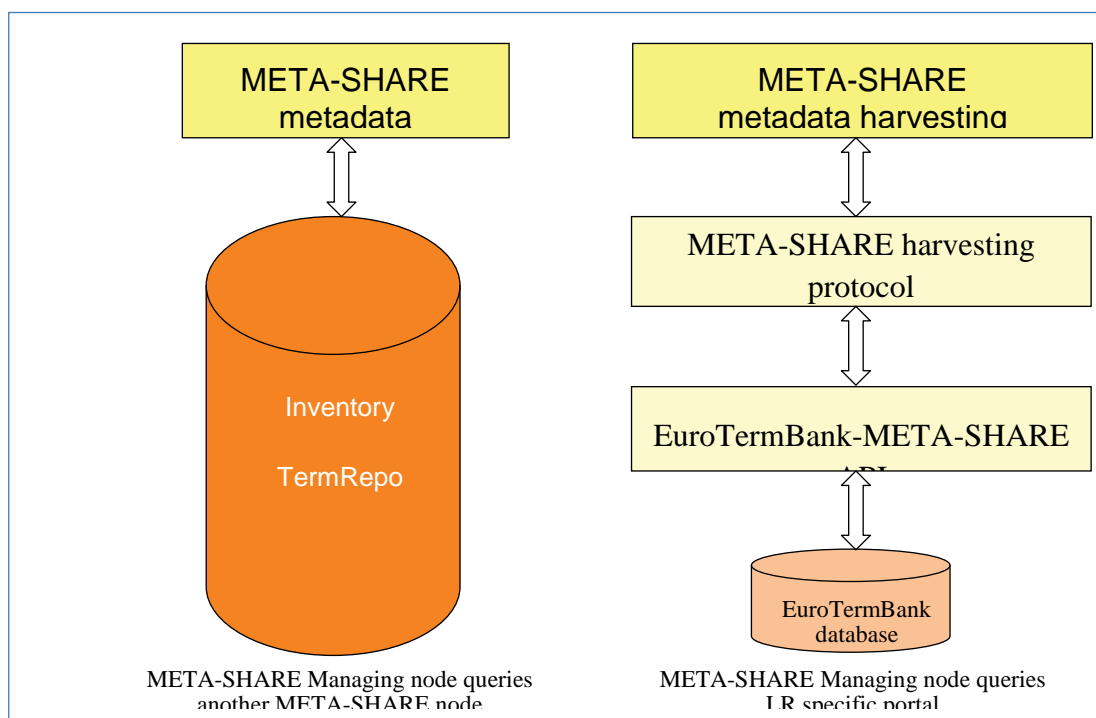

Figure 3 Repository harvesting: META-SHARE vs LR specific metadata harvesting

The development had two main components:

1. Implementation of the META-SHARE harvesting protocol

The META-SHARE harvesting protocol is published and available publicly at [2]

This is implemented in the following steps:

- **Authentication**: the client authenticates on the server;
- **Inventory request**: The client requests an LR inventory from the server;
- **Inventory response**: The server responds with its LR inventory.

For every server LR which is different or missing in the LR storage of the client:

- **LR request**: the client requests the LR from the server;
- **LR response**: the server responds with the requested LR;

  The client updates its LR storage with the server LR and takes note of where the LR came from;

  For every LR in the LR storage of the client which originally came from the server but which is not (anymore) in the LR inventory of the server, the client deletes the LR from its LR storage.

For flexibility and future scalability it is implemented as a web service in .NET Framework

2. Implementation of the EuroTermBank-META-SHARE API

This component serializes the data of a Language Repository and returns it in META-SHARE specific format upon harvesting request. Two web service methods do the job:

MetaShare_getCollectionList – returns the list of available resources – their IDs, creation and modification dates, and revision number. Based on this data META-SHARE takes a decision as to which resource descriptions to download, and which ones not. META-SHARE performs metadata harvesting on regular intervals (default – once an hour) and downloads only those resources that may have changed since last harvesting session.

MetaShare_getCollection – per supplied resource ID. This method generates a META-SHARE compliant metadata description XML and returns it to the calling process. The method communicates with the language resource specific database catalogue, and generates the metadata description XML based on internal data unique to each individual resource, like resource name, names of authors, size, contained languages, etc., and supplements it with generic metadata specific to the entire language resource specific portal, like portal name, address, contact person, etc.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<resourceInfo xsi:schemaLocation="http://www.ilsp.gr/META-XMLSchema http://metashare.ilsp.gr/META-
XMLSchema/v3.0/META-SHARE-Resource.xsd" xmlns="http://www.ilsp.gr/META-XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <identificationInfo>
    <resourceName lang="en">Athletics dictionary</resourceName>
    <description lang="en">Athletics dictionary is one of the terminology resources in EuroTermBank.
EuroTermBank is a centralized online terminology database for languages of new EU member countries
interlinked to other terminology resources. EuroTermBank platform provides a single access point to more
than 2 million terms in 27 languages
    </description>
    <url>http://www.eurotermbank.com/Collection.aspx?collectionid=209</url>
    <metaShareId>
                NOT_DEFINED_FOR_V2
    </metaShareId>
  </identificationInfo>
  <distributionInfo>
```

```xml
      <availability>available-unrestrictedUse</availability>
      <licenceInfo>
        <licence>proprietary</licence>
        <distributionAccessMedium>accessibleThroughInterface</distributionAccessMedium>
        <distributionRightsHolder>
          <organizationInfo>
            <organizationName lang="ka">Tilde</organizationName>
            <communicationInfo>
              <email>roberts.rozis@tilde.lv</email>
            </communicationInfo>
          </organizationInfo>
        </distributionRightsHolder>
      </licenceInfo>
      <iprHolder>
        <personInfo>
          <surname lang="en">Rozis</surname>
          <givenName lang="en">Roberts</givenName>
          <communicationInfo>
            <email>roberts.rozis@tilde.lv</email>
          </communicationInfo>
          <affiliation>
            <organizationName lang="en">Tilde</organizationName>
            <communicationInfo>
              <email>roberts.rozis@tilde.lv</email>
            </communicationInfo>
          </affiliation>
        </personInfo>
      </iprHolder>
    </distributionInfo>
    <contactPerson>
      <surname lang="lv">Rozis</surname>
      <givenName lang="lv">Roberts</givenName>
      <sex>male</sex>
      <communicationInfo>
        <email>roberts.rozis@tilde.lv</email>
        <address>Vienibas gatve 75a</address>
        <zipCode>LV 1004</zipCode>
        <city>Riga</city>
        <country>Latvia</country>
        <telephoneNumber>+371-67605001</telephoneNumber>
      </communicationInfo>
      <affiliation>
        <organizationName lang="en">Tilde</organizationName>
        <communicationInfo>
          <email>tilde@tilde.lv</email>
          <url>http://www.tilde.com</url>
          <address>Vienibas gatve 75a</address>
          <zipCode>LV 1004</zipCode>
          <city>Riga</city>
          <country>Latvia</country>
          <telephoneNumber>+371-67605001</telephoneNumber>
          <faxNumber>+371-67605750</faxNumber>
        </communicationInfo>
      </affiliation>
    </contactPerson>
    <metadataInfo>
      <metadataCreationDate>2011-12-31</metadataCreationDate>
      <metadataCreator>
        <surname lang="lv">Rozis</surname>
        <givenName lang="lv">Roberts</givenName>
        <sex>male</sex>
        <communicationInfo>
          <email>roberts.rozis@tilde.lv</email>
          <address>Vienibas gatve 75a</address>
```

```xml
            <zipCode>LV 1004</zipCode>
            <city>Riga</city>
            <country>Latvia</country>
            <telephoneNumber>+371-67605001</telephoneNumber>
        </communicationInfo>
        <affiliation>
            <organizationName lang="en">Tilde</organizationName>
            <communicationInfo>
                <email>tilde@tilde.lv</email>
                <url>http://www.tilde.com</url>
                <address>Vienibas gatve 75a</address>
                <zipCode>LV 1004</zipCode>
                <city>Riga</city>
                <country>Latvia</country>
                <telephoneNumber>+371-67605001</telephoneNumber>
                <faxNumber>+371-67605750</faxNumber>
            </communicationInfo>
        </affiliation>
    </metadataCreator>
    <metadataLastDateUpdated>2012-09-20</metadataLastDateUpdated>
</metadataInfo>
<usageInfo>
    <foreseenUseInfo>
        <foreseenUse>humanUse</foreseenUse>
    </foreseenUseInfo>
    <foreseenUseInfo>
        <foreseenUse>nlpApplications</foreseenUse>
        <useNLPSpecific>informationRetrieval</useNLPSpecific>
    </foreseenUseInfo>
    <actualUseInfo>
        <actualUse>nlpApplications</actualUse>
    </actualUseInfo>
    <actualUseInfo>
        <actualUse>humanUse</actualUse>
    </actualUseInfo>
</usageInfo>
<resourceDocumentationInfo>
    <documentation>
        <documentInfo>
            <documentType>book</documentType>
            <title lang="en">Towards Consolidation of European Terminology Resources: Experience and
Recommendations from EuroTermBank Project.</title>
            <author>Rirdance S., Vasiljevs A. (Eds.)</author>
            <editor>Edited by Signe Rirdance and Andrejs Vasi&#316;jevs</editor>
            <year>2006</year>
            <bookTitle>Towards Consolidation of European Terminology Resources: Experience and Recommendations
from EuroTermBank Project.</bookTitle>

    <url>http://www.eurotermbank.com/download/EuroTermBank_Towards_Consolidation_of_European_Terminology_Res
ources.pdf</url>
            <ISBN>9984-9133-4-1</ISBN>
            <keywords>terminology concept term harmonization classification EuroTermBank termbank</keywords>
            <documentLanguageName>English</documentLanguageName>
            <documentLanguageId>en</documentLanguageId>
        </documentInfo>
    </documentation>
</resourceDocumentationInfo>
<resourceComponentType>
    <lexicalConceptualResourceInfo>
        <resourceType>lexicalConceptualResource</resourceType>
        <lexicalConceptualResourceType>terminologicalResource</lexicalConceptualResourceType>
        <lexicalConceptualResourceEncodingInfo>
            <encodingLevel>other</encodingLevel>
            <linguisticInformation>lemma</linguisticInformation>
            <linguisticInformation>semantics-Domain</linguisticInformation>
```

```xml
        <linguisticInformation>semantics-Relations-Synonyms</linguisticInformation>
        <linguisticInformation>translationEquivalent</linguisticInformation>
        <conformanceToStandardsBestPractices>TBX</conformanceToStandardsBestPractices>
      </lexicalConceptualResourceEncodingInfo>
      <creationInfo>
        <creationMode>mixed</creationMode>
      </creationInfo>
      <lexicalConceptualResourceMediaType>
        <lexicalConceptualResourceTextInfo>
          <mediaType>text</mediaType>
          <lingualityInfo>
            <lingualityType>multilingual</lingualityType>
          </lingualityInfo>
          <languageInfo>
            <languageId>DE</languageId>
            <languageName>German</languageName>
          </languageInfo>
          <languageInfo>
            <languageId>EN</languageId>
            <languageName>English</languageName>
          </languageInfo>
          <languageInfo>
            <languageId>LV</languageId>
            <languageName>Latvian</languageName>
          </languageInfo>
          <languageInfo>
            <languageId>RU</languageId>
            <languageName>Russian</languageName>
          </languageInfo>
          <modalityInfo>
            <modalityType>writtenLanguage</modalityType>
          </modalityInfo>
          <sizeInfo>
            <size>665</size>
            <sizeUnit>entries</sizeUnit>
          </sizeInfo>
          <characterEncodingInfo>
            <characterEncoding>UTF-8</characterEncoding>
          </characterEncodingInfo>
          <domainInfo>
            <domain>social affairs</domain>
            <sizePerDomain>
              <size>665</size>
              <sizeUnit>entries</sizeUnit>
            </sizePerDomain>
            <conformanceToClassificationScheme>other</conformanceToClassificationScheme>
          </domainInfo>
          <domainInfo>
            <domain>miscellaneous industries</domain>
            <sizePerDomain>
              <size>665</size>
              <sizeUnit>entries</sizeUnit>
            </sizePerDomain>
            <conformanceToClassificationScheme>other</conformanceToClassificationScheme>
          </domainInfo>
        </lexicalConceptualResourceTextInfo>
      </lexicalConceptualResourceMediaType>
    </lexicalConceptualResourceInfo>
  </resourceComponentType>
</resourceInfo>
```

Code Sample 1. Metadata sample generated at LR-specific node

Behind the scenes: the language resource specific node must ensure additional fields in the database of the portal, and their proper handling in the life cycle of the resources – generation and storing of the unique ID, creation and modification date and revision number.

The implementation of the integration of a Language Resource specific node with META-SHARE is linked to META-SHARE via proxying: it connects to a META-SHARE Managing node just like any other META-SHARE node – LR metadata provider is proxied.

The integration of a Language Resource specific node with META-SHARE allows the user to access a specific resource located on a remote LR portal directly via a link supplied to META-SHARE.
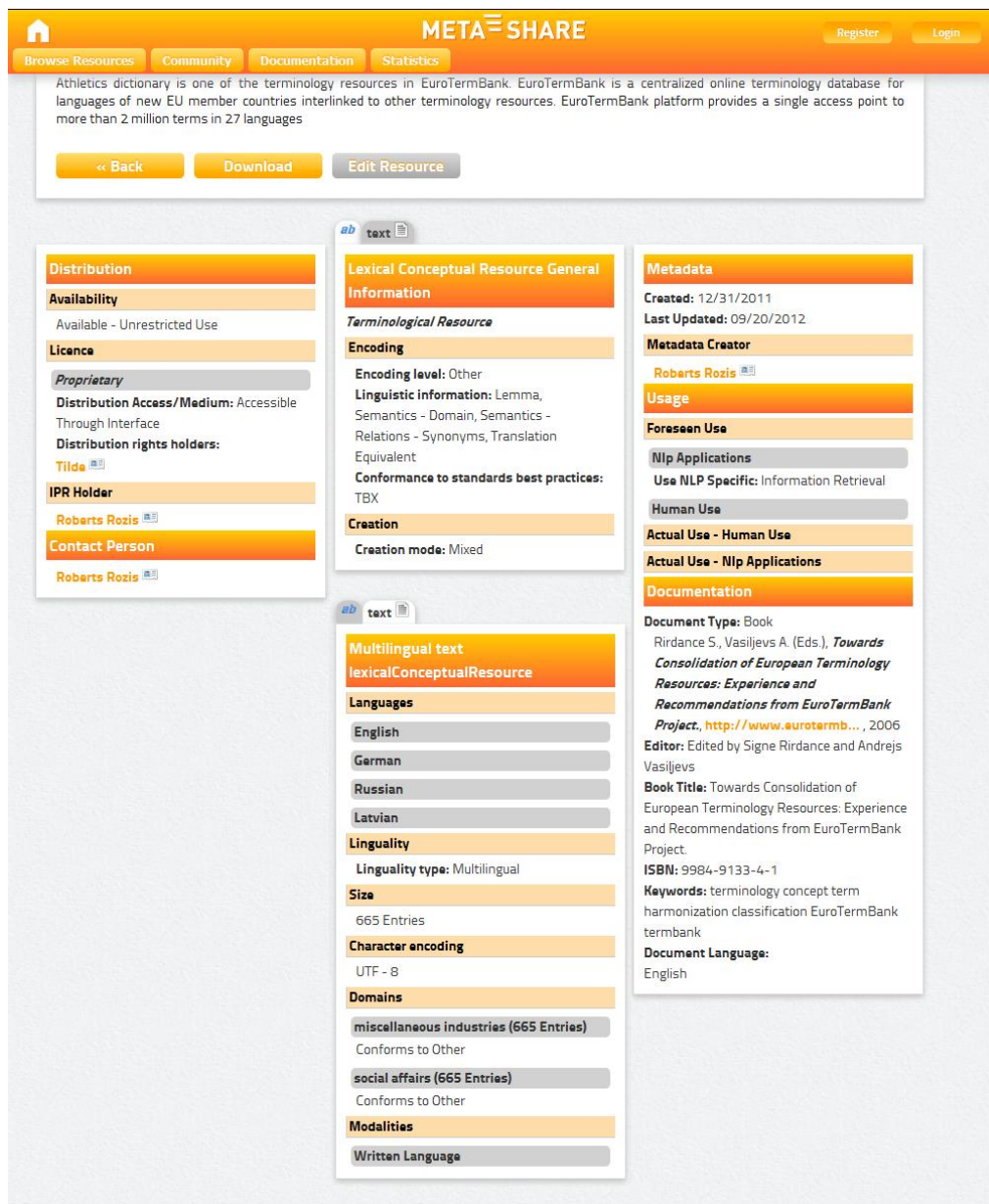

Figure 4 A record from EuroTermBank linked to and listed in META-SHARE

The interlinking yielded 99 additional terminology resources listed in META-SHARE. These resources are available in EuroTermBank for lookup search.

3. Synergies

Synergy within META-NORD

Proposals have been made to attempt to re-use Integration of a LR specific resource to other portals of LR specific resources like treebanks or wordnets. In particular, there was interest to link META-SHARE with the INESS project. Since the development of LR specific metadata harvesting was completed timely in M24 of META-NORD, the interlinking with INESS will be carried out shortly after end of the META-NORD project.

Synergies with CESAR and METANET4U

The invitation to collaboration was extended to the CESAR and T4ME projects with regard to collaboration in terminology identification and other means of terminology distribution. It is common practice to distribute terminology collections by selling them, and Tilde was suggesting another means of terminology distribution through online lookup in EuroTermBank. This alternative approach would not affect the market of terminology resources in its existing form as the terms are not given away and the data is protected from bulk copying. This approach promotes accessibility of terminology by its consumers – students, freelance translators, technical writers and researchers who only need access to terminology periodically. Unfortunately Tilde's invitation to collaboration was left without notice.

Conclusions

The Horizontal Action on multilingual terminology in Task 3.6 of META-NORD has resulted in a success in several areas as several term banks and resources have been identified, licensed, added to META-SHARE and integrated in the common terminology infrastructure of EuroTermBank.

A successful demonstration of creating an interlinked term base with META-SHARE has been achieved in Task 3.6, too. It is planned that the achievements from this Task of META-NORD will be further reused in other language resource specific portals to provide live and better exposition of their resources in the META-SHARE network.

Although some of the terminology resources were not included in META-SHARE as initially planned, some other resources have become identified and added to the scope of the project, and will be followed up even after the end of the META-NORD project.

Various parties involved in terminology work in Europe have different approach towards distribution and sharing of terminology data. The results of META-NORD show an open approach towards free sharing of terminology data. We anticipate that this will serve as a good example for other holders of data to share the terminology data freely. The interlinking of language resource portals will be yet another powerful tool which will result in many portals and live resources to be connected and available in META-SHARE thus ensuring its viability.

References

[1] CHAT 2012 Workshop
http://www.tilde.eu/tilde-research/workshop-creation-harmonization-and-application-terminology-resources

[2] META-SHARE Harvesting Protocol Specification
https://github.com/metashare/META-SHARE/wiki/META-SHARE-Harvesting-Protocol-v1.0