



NeOn: Lifecycle Support for Networked Ontologies

Integrated Project (IST-2005-027595)

Priority: IST-2004-2.4.7 – “Semantic-based knowledge and content systems”

D 2.4.1 Multilingual ontology support

Deliverable Co-ordinator: Elena Montiel-Ponsoda (UPM); Wim Peters (USFD); Guadalupe Aguado de Cea (UPM); Mauricio Espinoza (UPM)

Deliverable Co-ordinating Institution: Universidad Politécnica de Madrid (UPM); University of Sheffield (USFD)

Other Authors: Mari Carmen Suárez-Figueroa (UPM); José Ángel Ramos Gargantilla (UPM); Asunción Gómez-Pérez (UPM); Inmaculada Álvarez-de-Mon (UPM); Margherita Sini (FAO); Aldo Gangemi (CNR); Óscar Muñoz (UMP); Raúl Palma (UPM)

This deliverable aims at the provision of multilingual support for ontology development in terms of representation of multilingual information. The main contributions included in this document are: an overview of methods, techniques and tools used for the localizing process of some lexical and ontological resources; an analysis of the different ways for representing multilinguality at the various levels of a Knowledge Representation Base; an evaluation of the NeOn user requirements regarding multilinguality; and, finally, a proposal of a Multilingual Ontology Meta-model (MOM) for NeOn.

Document Identifier:	NEON/2007/D2.4.1/v1.15	Date due:	August 31, 2007
Class Deliverable:	NEON EU-IST-2005-027595	Submission date:	August 31, 2007
Project start date:	March 1, 2006	Version:	v1.15
Project duration:	4 years	State:	Final
		Distribution:	Public

NeOn Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities, grant number IST-2005-027595. The following partners are involved in the project:

<p>Open University (OU) – Coordinator Knowledge Media Institute – KMi Berrill Building, Walton Hall Milton Keynes, MK7 6AA United Kingdom Contact person: Martin Dzbor, Enrico Motta E-mail address: {m.dzbor, e.motta} @open.ac.uk</p>	<p>Universität Karlsruhe – TH (UKARL) Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB Englerstrasse 28 D-76128 Karlsruhe, Germany Contact person: Peter Haase E-mail address: pha@aifb.uni-karlsruhe.de</p>
<p>Universidad Politécnica de Madrid (UPM) Campus de Montegancedo 28660 Boadilla del Monte Spain Contact person: Asunción Gómez Pérez E-mail address: asun@fi.upm.es</p>	<p>Software AG (SAG) Uhlandstrasse 12 64297 Darmstadt Germany Contact person: Walter Waterfeld E-mail address: walter.waterfeld@softwareag.com</p>
<p>Intelligent Software Components S.A. (ISOCO) Calle de Pedro de Valdivia 10 28006 Madrid Spain Contact person: Jesús Contreras E-mail address: jcontreras@isoco.com</p>	<p>Institut 'Jožef Stefan' (JSI) Jamova 39 SI-1000 Ljubljana Slovenia Contact person: Marko Grobelnik E-mail address: marko.grobelnik@ijs.si</p>
<p>Institut National de Recherche en Informatique et en Automatique (INRIA) ZIRST – 655 avenue de l'Europe Montbonnot Saint Martin 38334 Saint-Ismier France Contact person: Jérôme Euzenat E-mail address: jerome.euzenat@inrialpes.fr</p>	<p>University of Sheffield (USFD) Dept. of Computer Science Regent Court 211 Portobello street S14DP Sheffield United Kingdom Contact person: Hamish Cunningham E-mail address: hamish@dcs.shef.ac.uk</p>
<p>Universität Koblenz-Landau (UKO-LD) Universitätsstrasse 1 56070 Koblenz Germany Contact person: Steffen Staab E-mail address: staab@uni-koblenz.de</p>	<p>Consiglio Nazionale delle Ricerche (CNR) Institute of cognitive sciences and technologies Via S. Martino della Battaglia, 44 - 00185 Roma-Lazio, Italy Contact person: Aldo Gangemi E-mail address: aldo.gangemi@istc.cnr.it</p>
<p>Ontoprise GmbH. (ONTO) Amalienbadstr. 36 (Raumfabrik 29) 76227 Karlsruhe Germany Contact person: Jürgen Angele E-mail address: angele@ontoprise.de</p>	<p>Food and Agriculture Organization of the United Nations (FAO) Viale delle Terme di Caracalla 1 00100 Rome Italy Contact person: Marta Iglesias E-mail address: marta.iglesias@fao.org</p>
<p>Atos Origin S.A. (ATOS) Calle de Albarracín, 25 28037 Madrid Spain Contact person: Tomás Pariente Lobo E-mail address: tomas.pariantelobo@atosorigin.com</p>	<p>Laboratorios KIN, S.A. (KIN) C/Ciudad de Granada, 123 08018 Barcelona Spain Contact person: Antonio López E-mail address: alopez@kin.es</p>

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

UPM

USFD

FAO

CNR

Change Log

Version	Date	Amended by	Changes
0.1	17-07-06	E. Montiel; G. Aguado	Initial draft: Termonography approach
0.2	18-08-06	E. Montiel; M.C. Suárez-Figueroa	Revision of criteria for comparing localization approaches
0.3	23-08-06	E. Montiel ; A. Gómez-Pérez	Inclusion of Systems of representation of multilingual information
0.4	20-09-06	E. Montiel; G. Aguado; I. Álvarez de Mon	Changes on definitions of lexical resources and on some content aspects
0.5	21-09-06	E. Montiel ; A. Gómez-Pérez	General content revision; changes on Systems of representation of multilingual information
0.6	26-09-06	E. Montiel; Margherita Sini; Ingrid Aldritt	AGROVOC and FAOTERM review
0.7	21-10-06	E. Montiel; G. Aguado; J.A. Ramos	Initial draft section 13. Representation of Multilinguality in NeOn
0.8	16-11-06	E. Montiel; J.A. Ramos; M.C. Suárez-Figueroa	Inclusión of WPs requirements, and advantages and disadvantages of representation systems for multilinguality
0.9	13-12-06	E. Montiel; W. Peters	Revision of multilingual information representation
1.0	19-12-06	E. Montiel ; W. Peters ; C. Caracciolo	Revision requirements for multilinguality
1.1	12-01-07	E. Montiel; W. Peters; J.A. Ramos	Revision figures of models and meta-models
1.2	10-05-07	E.Montiel; W. Peters; G. Aguado	□ inclusión of localization standards
1.3	21-05-07	E.Montiel	Revision of localizing approaches and inclusion of OntoLing and OncoTerm
1.4	11-06-07	E.Montiel; O. Muñoz; W. Peters	1 st version of Multilingual Ontology Meta-model
1.5	18-07-07	E. Montiel; W. Peters; G. Aguado	2 nd version of Multilingual Ontology Meta-model
1.6	26-07-07	E. Montiel; R. Palma; M.C. Suárez-Figueroa	Revision of OMV extension
1.7	07-08-07	G. Aguado; E. Montiel	Revision of the document
1.8	08-08-07	E.Montiel	Revision of Secion 11: addition of conclusions
1.9	08-08-07	W.Peters	Revision of chapters 11 and 12; addition of XLIFF

1.10	13-08-07	W.Peters	Revision of chapters 13, 14 and 15
1.11	14-08-07	M.Espinoza	Inclusion of section 17 (1 st prototype of the MOM)
1.12	15-08-07	E. Montiel; G. Aguado	Revision of sections 16 and 17
1.13	16-08-07	E.Montiel; M. Espinoza; W.Peters	Revisión of sections 16 and 17
1.14	08-10-07	E.Montiel	Revision of the whole deliverable
1.15	15-10-07	A.Tumilowicz	Final QA

Executive Summary

The first part of this deliverable aims at giving an overview of some methods, techniques and tools which are currently used for translating lexical on-line resources (LRs) (glossaries, dictionaries, databases, thesauri, lexicons) in the linguistic field and which could be, or are being, reused in the task of ontology localization. At the same time, we analyze current ontology localization methods with the aim of describing the SoA on multilingual resources and refining the list on NeOn multilinguality requirements.

In analysing the use of methods, techniques and tools for the localization or translation of the above mentioned types of LRs, we aim at describing strategies or steps in the translation task, which could be reused or adapted in the creation of multilingual ontologies. We will begin with the description of those LRs that contain less semantic information and have a poorer internal structure, to end up with an analysis of LRs considered more complex and elaborated in their structure and content. Finally, we will survey current ontology localization strategies, which share some of these strategies with LRs or even nourish from them for the translation task.

The research is limited to multilingual LRs which are representative examples, have an authority within the linguistic and applied science communities, and are supported by relevant research groups (joint projects of universities and private companies, indicated in the report), international organisms (e.g. EU, FAO) or national institutions.

LRs are grouped in the following clusters (sections 5 to 11):

- Glossary localization approaches.- **FAOTERM**
- Database localization approaches. - **FishBase**
- Dictionary localization approaches.- **Eurodicautom**
- Thesauri localization approaches.- **Agrovoc, Eurovoc**
- Lexicon localization approaches.- **EuroWordNet**
- Ontology localization approaches.- **Thermontography, LabelTranslator, GENOMA, OncoTerm, OntoLing**

After the initial survey, this deliverable is devoted to the analysis of the different ways for representing multilinguality at the various levels of a Knowledge Representation Base (section 12 ff.). The differentiated levels are:

- 1) Interface level
- 2) Metadata level: OMV

- 3) Knowledge Representation level
- 4) Data level

For our modelling purposes we stay within the confinements of an ontology and concentrate on the first three levels. In this sense, the first step has been an evaluation of the main requirements regarding multilinguality that have been listed in the different NeOn WPs. Those requirements impose some restrictions in the final proposals for representing multilinguality in NeOn. Bearing in mind the main implications derived from the different WP requirements, we propose meta-models and models for the representation of multilingual data at the identified levels. To each proposal we attach an example explaining its convenience.

The proposed Multilingual Ontology Meta-model (MOM) for NeOn is finally described. However, and for the time being (M18), this model is not going to be implemented in the 1st prototype of the NeOn toolkit. Until the definitive MOM is produced, a 1st prototype is to be implemented based on the OWL ontology meta-model and a supporting tool for the translation of ontology labels called LabelTranslator. The first MOM Prototype as well as the LabelTranslator functionalities and architecture are as well explained.

Table of Contents

NeOn Consortium	2
Work package participants	3
Change Log	3
Executive Summary	4
Table of Contents.....	6
List of tables.....	9
List of figures	10
1. Introduction	12
2. Interoperability with Knowledge Representation Standards	14
2.1 TMF	15
2.2 LMF	16
2.3 SKOS Core	17
3. Lexical resources (LRs)	17
4. Evaluation framework used to compare the lexical resources on each section	21
5. Glossary localization approaches	22
5.1 FAOTERM.....	22
5.1.1 Short description of FAOTERM	22
5.1.2 Comparison of FAOTERM against the evaluation framework.....	22
6. Database localization approaches	27
6.1 FishBase	27
6.1.1 Short description of the FishBase.....	27
6.1.2 Comparison of FishBase against the evaluation framework	27
7. Dictionary localization approaches.....	32
7.1 Eurodicautom	32
7.1.1 Short description of Eurodicautom	32
7.1.2 Comparison of Eurodicautom against the evaluation framework.....	32
8. Thesauri localization approaches	36
8.1 AGROVOC.....	36
8.1.1 Short description of AGROVOC	36
8.1.2 Comparison of AGROVOC against the evaluation framework.....	36
8.2 Eurovoc.....	40
8.2.1 Short description of Eurovoc.....	40
8.2.2 Comparison of Eurovoc against the evaluation framework.....	40
9. Lexicon	46
9.1 EuroWordNet (EWN).....	46
9.1.1 Short description of EWN	46
9.1.2 Comparison of EWN against the evaluation framework.....	46
10. Ontology localization approaches	53
10.1 Termontography approach.....	53

10.1.1	Short description of the Termontography approach	53
10.1.2	Comparison of the localization approach against the evaluation framework	53
10.2	LabelTranslator approach	56
10.2.1	Short description of the LabelTranslator approach	56
10.2.2	Comparison of the LabelTranslator against the evaluation framework	57
10.3	OntoLing Tab approach	60
10.3.1	Short description of the OntoLing approach	60
10.3.2	Comparison of the OntoLing against the evaluation framework	60
10.4	GENOMA-KB approach	64
10.4.1	Short description of GENOMA-KB	64
10.4.2	Comparison of GENOMA-KB against the evaluation framework	64
10.5	OncoTerm approach	69
10.5.1	Short description of the OncoTerm approach	69
10.5.2	Comparison of the OncoTerm against the evaluation framework	69
11.	Conclusions and Summarizing Tables	74
11.1	Main Conclusions to the Multilingual Resources Survey	74
11.1.1	Localization Approaches of LRs	74
11.1.2	Localization Approaches of Ontologies	74
11.2	Summarizing tables	77
12.	Representation of multilinguality in NeOn	81
12.1	Introduction	81
12.2	Standardization of localization	81
12.2.1	TMX (Translation Memory Exchange)	82
12.2.2	XLIFF	84
12.2.3	MLIF	86
12.3	Requirements for multilinguality in NeOn and restrictions	89
WP1	89
WP6	90
WP7	90
WP8	92
Summary of the main implications of WP requirements	93	
12.4	Rationale for a three layered approach and evaluation criteria	93
12.4.1	Evaluation criteria for interface level	93
12.4.2	Evaluation criteria for metadata level	93
12.4.3	Evaluation criteria for KR level	94
13.	Representation of multilinguality at the Interface level	95
13.1	Multilingual interface at message level	95
13.2	Multilingual interface at content level	96
13.3	Advantages and disadvantages of a multilingual query	96
13.4	Advantages and disadvantages of adding a new language to the interface	96
14.	Multilinguality in a Knowledge Representation System (KRS)	97
14.1	OMV level: Modification/Extension of the Core Model	97
14.1.1	Advantages and disadvantages of both representation systems	98
14.2	KR level	99
14.2.1	Modified Ontology Meta-model	99
14.2.2	Advantages and disadvantages of a Modified Ontology Meta-model	101
14.2.3	Ontology Meta-model linked to a Linguistic Information Respository (LIR) Model	102
14.2.4	Advantages and disadvantages of an Ontology Meta-model linked to a LIR Model	103
15.	Ontology Models: realization and instantiation	104
15.1	1 st Proposal: Realization of the Modified Ontology Meta-model	104

15.2 2 nd Proposal: Realization of the Ontology Meta-model linked to a LIR Model	105
15.3 Hybrid systems.....	107
16. The Multilingual Ontology Meta-model proposed for NeOn	109
16.1 NeOn Ontology Meta-model linked to the LIR Model.....	109
16.1.1 <i>The choice of Model</i>	109
16.1.2 <i>Description of the classes:</i>	111
16.1.3 <i>Description of the relations:</i>	114
16.1.4 <i>LIR properties</i>	115
16.2 OMV extension for capturing multilinguality: LexOMV	116
17. 1st Prototype of the NeOn Multilingual Ontology Meta-model	118
17.1 Requirements specification	118
17.2 NeOn Multilingual Meta-model implementation proposal.....	119
17.3 Description of the 1 st Prototype of the NeOn Multilingual Meta-model.....	121
17.3.1 <i>Architecture</i>	122
17.3.2 <i>Ranking for ordering translations</i>	126
17.4 Use Cases of the 1 st Prototype.	130
17.4.1 <i>Use Case: Add Language</i>	130
17.4.2 <i>Use Case: Translate an ontology label</i>	132
18. Future work	136

List of tables

Table 1: Comparison of lexical resources	20
Table 2: FAOTERM Localization Tools.....	24
Table 3: Main Language and Translation Tools of the DGT	41
Table 4: Steps used for localizing the Dutch WordNet	48
Table 5: Steps used for localizing the Spanish WordNet.....	49
Table 6: Tools for supporting the process of term extraction and translation	54
Table 7: Steps, sources and techniques used for localizing in LabelTranslator	57
Table 8: Steps, sources and techniques used for localizing in OntoLing.....	61
Table 9: Steps, sources and techniques in the localization of GENOMA-KB	66
Table 10: Steps, sources and techniques in the localization of OncoTerm	70
Table 11: General description of approaches	77
Table 12: Aims, languages and domains involved in the resources	78
Table 13: Steps and tools used for localization	79
Table 14: Compulsory and Optional Attributes of the <header>	82
Table 15: Inline Elements	83
Table 16: TMX Attributes	83
Table 17: Criteria for identifying advantages and limitations of MOM.....	108

List of figures

Figure 1: TMF structural Skeleton.....	15
Figure 2: Relationship between term entries and language sections in TMF	15
Figure 3: TMF language model.....	16
Figure 4: LMF core.....	17
Figure 5: Lassila and McGuinness (2001) categorization	19
Figure 6: FAOTERM terminology workflow system (TRG/GICM 2006, provided by FAO).....	23
Figure 7: FAOTERM interface	25
Figure 8: Search result for the <i>Poecilia gillii</i> species known commonly as “molly” in Costa Rica...	29
Figure 9: Fragment of the Species table for <i>Rainbow trout</i> in Spanish.....	30
Figure 10: FishBase GLOSSARY	31
Figure 11: Eurodicautom interface.....	34
Figure 12: Results for a searched term in Eurodicautom.....	35
Figure 13: Interface of the AGROVOC Thesaurus, 1st step in the search	38
Figure 14: Interface of the AGROVOC Thesaurus, 2nd step in the search	38
Figure 15: AGROVOC Systems of representation of multilingual information	39
Figure 16: DGT translation workflow (DGT 2005).....	41
Figure 17: Results for the searched term “fish” in the 1 st step of the search	44
Figure 18: Results for the searched term “fish” in the 2 nd step of the search.....	44
Figure 19: Subject fields and microthesauri of the EC in Eurovoc.....	45
Figure 20: The global architecture of the EWN database (Vossen 2004).....	47
Figure 21: Building steps in EWN (Vossen 2002).....	50
Figure 22: Interface Meaning 2.0	51
Figure 23: General outline of two wordnets linked to the ILI (Vossen 2002)	52
Figure 24: Termontography workflow (Kerremans et al. 2004b).....	54
Figure 25: LabelTranslator interface	58
Figure 26: API structure (Gantner 2004).....	59
Figure 27: Selection of LRs in OntoLing	61
Figure 28: Linguistic Browser Panel in OntoLing	62
Figure 29: Inclusion of multilingual data in ontologies in OWL	63
Figure 30: Knowledge Base architecture (Feliu <i>et al.</i> 2002)	65
Figure 31: Hyperonymy relations for the term “cell” in GENOMA-KB	67
Figure 32: GENOMA-KB architecture support (Hospedales y Rodríguez 2004)	68
Figure 33: Results from the search of “air-contrast-x-ray” in the OncoTerm resource.....	72
Figure 34: Linguistic dimension of the OncoTerm resource	73
Figure 35: MLIF Metamodel.....	87
Figure 36: MLIF Metamodel with related Data Categories	87

Figure 37: Simultaneous multilingual interface messages.....	95
Figure 38: Alternately monolingual interface messages in a multilingual system	96
Figure 39: Example of a possible extension to the OMV	98
Figure 40: Tuple with multiple values about linguistic information in OMV	98
Figure 41: Classes and Properties of the OWL DL Meta-model for NeOn (D1.1.1: 25)	100
Figure 42: MOM represented by <code>Label</code> , <code>Definition/Gloss</code> and <code>Source</code> classes linked to <code>Class</code> and <code>Property</code>	101
Figure 43: Two Models: the Ontology Meta-model and the LIR Model.....	102
Figure 44: Example of a MOM represented by a LIR Model linked to the Ontology Meta-model .	103
Figure 45: Example of an Ontology Model based on a Modified Ontology Model with multilingual Instances associated to Classes.....	105
Figure 46: Example of an Ontology Model linked to Multilingual LIR Instances	106
Figure 47: The LIR model	110
Figure 48: Extension of the OMV Core to capture multilingual data: LexOMV	117
Figure 49: Three layer architecture of the 1 st prototype of the NeOn Multilingual Meta-model.....	120
Figure 50: Schema of the 2 nd prototype of the NeOn Multilingual Meta-model.....	121
Figure 51: Main components of the <i>LabelTranslator</i> plugin.....	122
Figure 52: A screenshot of the Ontology Navigator with the action “Translate”	123
Figure 53: A sample of linguistic information in the Entity Properties View	124
Figure 54: User dialog with the translation results of an ontology label	125
Figure 55: Main steps of the translation ranking method	127
Figure 56: GUI prototype of the language preferences.....	131
Figure 57: GUIs to translate an ontology label using LabelTranslator	134
Figure 58: AGROVOC planned revision workflow	138
Figure 59: AGROVOC management tool.....	138

1. Introduction

In this survey, our objective is to analyse different localization approaches in order to describe the steps which led to the creation of the existing multilingual lexical resources. In this document the terms *localize* and *translate* will be used with the same meaning. Therefore, we find it appropriate to define both concepts and determine the reason for a possible distinction between them.

To **localize** means literally “to make local” or “to orient locally” (Merriam-Webster Online Dictionary). In the Free Encyclopaedia Wikipedia we find it generally defined as “the adaptation of an object to a locality”. Localization can be applied to many domains. In economics, for example, localization is the way to “adapt products for non-native environments”. In web design and software, localization refers to “the adaptation of language, content and design to reflect local cultural sensitivities”.

The concept of **translation** has received much more attention throughout history as the activity of translating has been carried out since different language communities exist and communicate with each other. Following the functionalist approach to translation, it can be described as “a type of transfer where communicative verbal and non-verbal signs are transferred from one language into another.(...) Translation is thus an intentional, purposeful action that takes place in a given situation; ...” (Vermeer 1983, cited in Nord 1997). Functionalists put emphasis on the fact that every translation is intended to fulfil a specific **function** on a specific target culture, hence the name of their approach. Translation cannot be reduced to a one-to-one-word translation, but in every translation process there are many aspects that have to be taken into account. These are:

- Intention of the text – to inform, to convince, to give orders...
- Target-text addressee(s) – adults, children, experts, scientists...
- Time and place of the text reception – a company, a country, for one year, for a month...
- Medium over which the text will be transmitted – monolingual or bilingual web pages, brochures...
- Motive for the production or reception of the text – presentation of a new product, celebration of an anniversary...

However, the most important factor which has to be borne in mind is the **function of the translation**, i.e., the role the translation is going to play in the target culture.

- If the aim of the translation is to *document* the target reader about a situation in the original language and culture, reproducing the same intention, it may result in a text with a *foreign flair* for the target reader, so that he or she is conscious of the character of translation of the text.
- If the translation aims at producing in the target reader the same effect the original text produced in the original reader, the translator may have to adapt many aspects of the text, or even change or omit facts, so that the target reader feels the text as original of his or her culture.

Many practitioners and translator theorists agree about this difference and talk about *overt vs. covert translation* (House 1977), and *documentary vs. instrumental translation* (Nord 1989).

Notwithstanding, after having defined both concepts, we have to admit, that localization and translation are equivalent, when by translating we understand the second option considered, i.e., to “produce in the target reader the same effect the original text produced in the original reader”.

Some authors in web design domain, however, think localization is “a substantially more complex issue”¹ than translation, and restrict translation to the linguistic part of the process, without having into account that the concept of translation –or instrumental translation- considers both the fact of having to adapt the linguistic information and non-verbal aspects of all kinds. The main point here is that in the localizing industry the activity of translating is not limited to a text of 5.000 words as source material, but consists of software programs in which technical aspects play a decisive role. This is why some authors in the software localizing and web design domains think *localization* is “a substantially more complex issue”² than translation, and they restrict translation to the linguistic part of the process. However, and according to recent analysis on translation (cf. Hurtado Albir 2001:87ff) the activity of translating has become a more complex process, in which not only pure linguistic aspects are considered, but other abilities are demanded from translators, specifically computer abilities as, for example, the use of different text formats, translation supporting tools, or even image manipulation software.

In the description of the “Guidelines for building multilingual Web sites” from the EURESCOM Project³, the “typical localization process” was defined as follows: “The **localization process** is divided in three main stages: planning, translation and after translation” (from Esselink 2000: 17).

- **Planning** is one of the most decisive factors when undertaking a localization project and it consists in being able “to anticipate possible problems and try to find solutions to prevent them before they appear”. Planning manager, Project developers (in the origin and target country), Translators, Localization Engineers and Proof-reader are involved.
- **Translation** is the second stage of the process and represents the core of localization, “where real translation and adaptation to other languages take place”.
- **After Translation**, which is generally carried out by the translator, the first priority is to check whether all information has been properly translated into the target language. Other aspects of the translation and adaptation can be proofed together with the Proof-reader.

It is worth mentioning that Localization Specialists are also called by Esselink in his book (2000:16), Senior Translators or simply Translators, and that they are in charge of reviewing the work other translators do, setting standards and managing terminology. The author also explains that those linguists who translate software applications are called *localizers*, “because they get involved in other project activities such as software user interface resizing”. This means that translators are main actors in the localizing process, although they work together with localization engineers, CAT tools experts and other specialists, depending on the complexity of the project.

In a parallel way, the typical translation process could also be divided in three similar stages:

- **Translation brief and source text analysis** is the first and main stage in the translation process. In order to find out the purpose of the target text that will guide the translator throughout the process, he or she should compare the source text against the translation brief, if it exists, or against the client’s demands, in order to infer the intended text *function*, target text addressees, and motive for the production or reception of the text (Nord 1997: 60), among other relevant information. This analysis will determine all decisions the translator will have to take during the whole activity development.
- **Translation** is the central part of the translation process, and where the translator produces a *functional* target text, functional in the sense that “it meets the demands of the translation brief” (Nord 1999:21). During this time-enduring stage, other parallel tasks take place, as on-line research, glossary construction, etc. All of them require the help of a wide range of translation supporting tools.

¹ <http://www.w3.org/International/questions/qa-i18n.en.html>

² <http://www.w3.org/International/questions/qa-i18n.en.html>

³ <http://www.eurescom.de/>

- **Revision and proof-reading** is the last step in the translation process and should be done by the translator him- or herself, and by a proof-reader, who normally is a translation colleague. Both of them have to keep in mind the purpose of the translation brief and check sense, grammar and style.

After this brief survey it could be stated that the localizing activity, applied to software or ontologies, as in our case, could even be considered a new speciality in the translation field, a new form of translation. We support this statement after having found many parallelisms between the localization process and the translation one. Obviously, we are aware of the technical limitations of translators, but it is also evident that the translation activity has gone through many stages in history, has adapted to the new technologies and forms of required translations, and this time it will not be an exception. Universities and other higher education institutions will have to adapt to the new times to form translators according to the current needs. However, and for the time being, we consider that the use of the concept *localization* in the Computer Science domain, and more specifically in the Knowledge Engineering field, describes very precisely the wide range of activities involved in this task and the high number of actors interoperating in the process, which have not been identified until now in the translation process. In this sense, we could say that a *localizer* is a **translator specialized in the translation and adaptation of software and web products**. Following this line of thought, we agree to define *ontology localization* as **all steps carried out in the process of adapting an ontology to a concrete language and culture community**. However, and for the purposes of this work, localization and translation are sometimes used interchangeably.

After this introduction about the concepts of localization and translation, in section 2 we will define the different kinds of lexical resources dealt with in this survey. This will represent the starting point for the subsequent analysis of the specific resources that are to be taken into account. We also briefly compare the different types of lexical resources, establishing as the main criterion the semantic information they provide. In section 3, the main set of criteria used to analyze the different localization approaches are presented. In the next sections, Sections 5, 6, 7, 8, and 9, we provide authoritative examples of each type of lexical resource, and we compare them following the evaluation framework previously established. In section 10 we do the same with ontology localization approaches. section 11 of this survey presents the conclusions of the multilingual resources survey, and 3 Summarizing Tables conclude the first part of this research. With section 12, the second part of the Deliverable starts. As already mentioned in the Executive Summary, the second part is devoted to the analysis of the possible models and methods for representing Multilinguality in Knowledge Based Systems, and the presentation of the Multilingual Ontology Meta-model proposed for NeOn.

2. Interoperability with Knowledge Representation Standards

When representing multilinguality in ontologies, it is important to take into account several standardization initiatives within the fields of linguistics and terminology.

The potential integration of terminological and linguistic knowledge bases into the NeOn model requires interoperability with existing and proposed standards for the representation and integration of terminological and linguistic knowledge. This integration supports knowledge exchange between heterogeneous sources, and mappings between them provide assistance with re-engineering activities.

The existing standardization efforts taken into account are listed below:

ISO 16642:2003, *Computer applications in terminology – TMF (Terminological Markup Framework)*

Described in section 2.1

ISO 24613 *Language Resource Management – LMF: Lexical Markup Framework*

Described in section 2.2

ISO 12620, *Terminology and other language resources: Data categories*.

Data categories are linguistic/terminological notions such as Term and PartofSpeech, which are used in the framework models above.

ISO 639-2:1998, ISO DIS 639-3:2005:

Both codes for the representation of languages.

TMF and LMF are briefly described below.

2.1 TMF

The TMF framework⁴ (and the associated TermBase eXchange format; TBX⁵) captures the underlying structure and representation of computerized terminologies. Its overall architecture is illustrated in Figure 1.

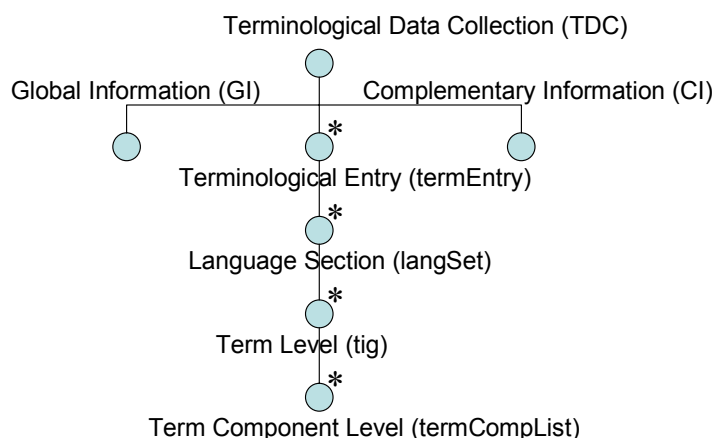


Figure 1: TMF structural Skeleton

Multilingual information at ontology resource level in this framework is positioned under Global Information.

Multilingual information at ontology element level is contained in the Language Section in a term entry. Each term entry may contain more than one language section (see Figure 2).

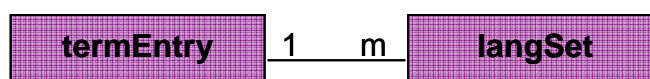


Figure 2: Relationship between term entries and language sections in TMF

⁴ <http://www.loria.fr/projets/TMF/>

⁵ <http://www.lisa.org/standards/tbx/>

The XML representation example below shows lexicalizations in two different languages for a particular concept.

```
<termEntry id='ID67'>
  <definition='a type of flatfish'>
  <langSet lang='en'>
    <tig>
      <term>plaice</term>
      <termNote type='termType'>fullForm</termNote>
    </tig>
  </langSet>
  <langSet lang='nl'>
    <tig>
      <term>schol</term>
    </tig>
  </langSet>
</termEntry>
```

Graphically, this looks as follows:

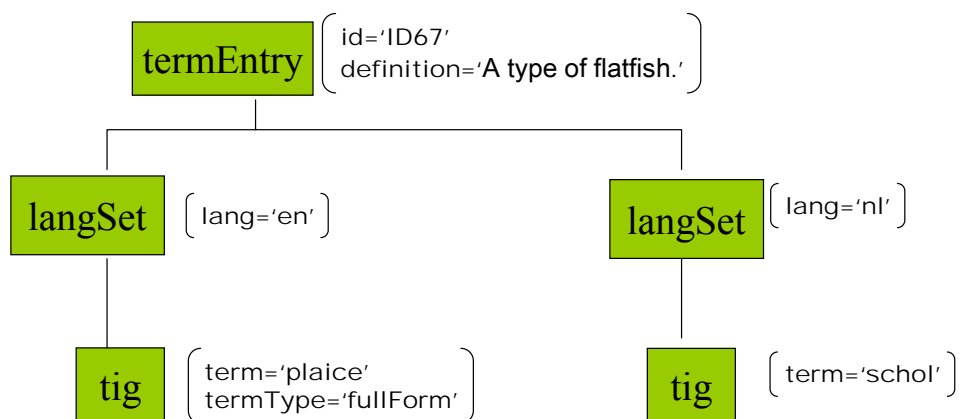


Figure 3: TMF language model

2.2 LMF

The Lexical Markup Framework (LMF; ISO/CD 24613) is an abstract meta-model that provides a common, standardized framework for the construction of computational lexicons. The LMF ensures the encoding of linguistic information in a way that enables reusability in different applications and for different tasks. The LMF provides a common, shared representation of lexical objects, including morphological, syntactic, and semantic aspects. It is under development and expected to be defined a standard in 2007.

The core model comprises a meta-model, i.e., the structural skeleton of the LMF, which describes the basic hierarchy of information included in a lexical entry. It revolves around the data categories Lexical Entry, which is constituted by a combination of Form and Sense (see Figure 4).

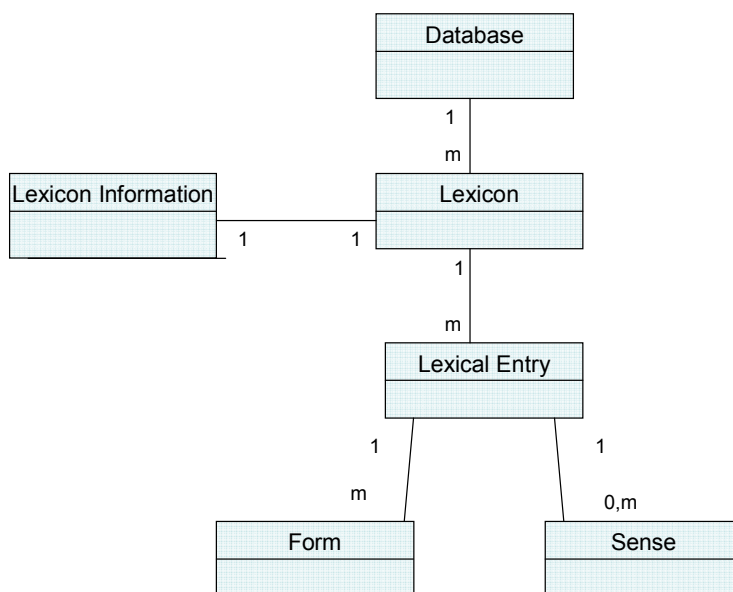


Figure 4: LMF core

The language codes are specified in ISO 639 and the country codes in ISO 3166.

Information types from ISO standards will be re-used in various cases below.

2.3 SKOS Core

SKOS Core⁶ (Simple Knowledge Organization Systems) provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies.

The practical goal of SKOS Core is to support the interoperation of software systems via a common data language. It is envisaged that it will be extended with modules that specifically model more fine-grained information.

At the moment, Skos core covers the following data objects for handling labels:

- `prefLabel`: a preferred label
- `altLabel`: an alternative label
- `hiddenLabel`: a hidden label (not exposed to any search methods)

3. Lexical resources (LRs)

In this report, we will analyse methods, techniques and tools currently used for the localization of some relevant lexical resources and ontologies. We have selected six types of such resources: **glossaries**, **databases**, **dictionaries**, **thesauri** and **lexicons**, as well as **ontologies**, that in a way can also be considered linguistic artefacts, although their main objective –as explained below– is in short the representation of the conceptual structure of a field of

⁶ <http://www.w3.org/2004/02/skos/core/>

knowledge. Although among the community of linguists and terminologists there is no absolute consensus on the definition of the different kinds of LRs, for most of them the difference depends on “the information they need to express and the richness of their internal structure” (Lassila & McGuinness 2001). However, the internal structure is not the only characteristic to take into account when defining LRs. Purpose of the resource, end users and representation of the information are other important features that have to be considered. Bearing all this in mind, we have distinguished the following types of LRs as they are the most representative, as well as the most widely accepted⁷. In all cases we talk about online resources.

- A **glossary** is a collection (...) of specialized terms with their meanings, according to the Merriam-Webster Online Dictionary⁸. In the *ontology engineering* literature we find the following definition “glossaries are lists of terms with their meanings specified as natural language statements” (Gómez-Pérez *et al.* 2003⁹).
- **Databases** are collections of records –or pieces of knowledge- stored in a computer in a systematic way, so that a computer program can consult it to answer questions. Records are usually organized as a set of data elements, for best retrieval and storing (Wikipedia).
- **Dictionaries** are, according to Wikipedia¹⁰, lists of words with their definitions in natural language. The headword -or main word- in the majority of the dictionaries is the lemma. Many dictionaries also provide pronunciation and grammatical information, derivations, etymologies, usage guidance and examples in phrases or sentences. Bilingual dictionaries also offer an explanation or translation of the headword in another language.
- According to the definition of the Merriam-Webster Online Dictionary, a **thesaurus**, when dealing with on-line resources in the domain of computer science is “a controlled and dynamic documentary language containing semantically and generically related terms, which comprehensively covers a specific domain of knowledge” or in a more specific way, “a controlled list of descriptors (preferred terms) and non-descriptors (non-preferred terms) related by semantic (that is, hierarchical, associative, or equivalence) links”.
- Although a **lexicon** is oft identified with a dictionary in general dictionaries as the Merriam-Webster Online Dictionary, or defined as the vocabulary of an individual speaker or group of speakers, in lexical semantics, a lexicon is considered a LR with abundant semantic and syntactic information, and richer internal structure. We have chosen the definition of lexicon within the framework of the Functional Lexematic Model (Martín Mingorance 1998, Faber & Mairal 1999), which is again based on the Dik’s Functional Grammar conception of lexicon (Dik 1978). Following those models, a lexicon is a “network of information about words and its contexts” (Faber and Mairal 1992: 63). The central unit of the lexicon is the word or lexeme, which is provided with its meaning definition, the grammatical information necessary for its use in different contexts, as well as morphology, phonology and part of speech”. A set of lexemes –called lexical domain- lexicalizes a determinate conceptual domain, which consists of different more specific sub-domains. Lexemes are organized primary in a hierarchical way, but additional hierarchical relations are also taken into account.

⁷ Our purpose is not to analyse and compare the existing definitions for the resources above mentioned so as to give a definitive definition, but to establish a framework for analysing lexical resources and justify their convenience.

⁸ <http://www.m-w.com/dictionary/>

⁹ <http://webode.dia.fi.upm.es/ontologicalengineering/>

¹⁰ <http://www.wikipedia.org/>

- The most quoted definition of **ontology** in Artificial Intelligence literature is the Gruber's one (1993). The author defined ontology as "an explicit specification of a conceptualization". Studer and colleagues (Studer *et al.* 1998: 185) based on this definition and the one by Borst (1997) which said that "Ontologies are defined as a formal specification of a shared conceptualization", and merged both to state that: "*Conceptualization* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. *Explicit* means that the type of concepts used, and the constraints on their use are explicitly defined. *Formal* refers to the fact that the ontology should be machine-readable. *Shared* reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted group".

After this brief review of the main characteristics of each resource, we will use Lassila and McGuinness (2001, cited in Gómez-Pérez *et al.* 2003:28) *ontology categorization criteria* to compare the above defined resources. Lassila and McGuinness classified ontologies according "to the richness of their internal structure and also to the subject of the conceptualization", and pointed out the following categories: controlled vocabularies, glossaries, thesauri, informal is-a hierarchies, formal is-a hierarchies, formal is-a hierarchies with instances, frames, ontologies with value restriction, and ontologies with general logical constrains, as Figure 5 shows. In compliance with those criteria, ontologies were classified in lightweight and heavyweight ontologies in a continuous line. Thus, those ontologies with less semantic information and a poorer internal structure were considered *lightweight ontologies* and were placed to the left of the crossing line; and those that were able to express a considerable quantity of semantic information and organize it following psycholinguistic principles received the attribute of *heavyweight ontologies*, and were placed to the right.

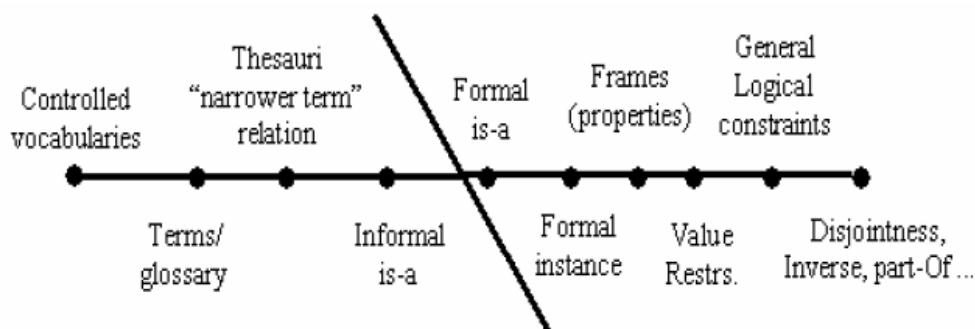


Figure 5: Lassila and McGuinness (2001) categorization

For the purposes of this research we will talk about **lightweight** and **heavyweight** resources. According to previous considerations, **Glossaries** are considered lightweight LR, i.e., the organization of a glossary follows an alphabetical order and the information contained is normally limited to the definition of the lexical items or the equivalent in another language at the most. **Databases** fall in the same group, since no semantic information is available. A **dictionary**, also regarded as a lightweight LR, is organized in an alphabetical order and provides not only a definition for the headword, but also grammatical information (part-of-speech, gender, number, etc.), semantics between lexical units (limited to usage guidance provided by examples and references to related terms), and some additional information as derivations and etymologies. Although currently most glossaries and dictionaries have an electronic version, they are traditionally founded in paper format. Concerning subject matter, it is possible to find both general and specialized glossaries, dictionaries and databases.

Thesauri are lists of words and phrases close in meaning to each other and are organized following the principle of semantic locality. It is supposed that when you look a word in a thesaurus you already know its meaning and for that reason, definitions are not compulsory in a thesaurus.

However, semantic relationships between terms (hierarchical, associative, or equivalence relationships) are the main feature of this sort of LR. Thesauri usually deal with a specific domain and can be found in both electronic and paper format. The same applies for **lexicons**. Lexicons also organize words and expressions depending on semantic relations, and unlike thesauri, they do “supply explicit hierarchy” (Gómez-Pérez *et al.* 2003) and additional kinds of semantic information as antonymy or meronymy.

Finally, **ontologies** differ from **lexicons** and **thesauri** in that all concepts in an ontology are organized hierarchically around a unique concept, superordinated to the rest, and, more important, that relations between concepts are more specific than in the other resources and capture consensual knowledge, i.e. semantics of the domain are shared and accepted by experts. Moreover, information stored in ontologies can be interpreted not only by humans, as in the case of lexicons and thesauri, but also by machines (Arano 2005). Although dealing with the different types of ontologies that exist goes beyond the purpose of our analysis, and would alone be the subject of a survey (see Gómez-Pérez *et al.* 2003: 26-37 for this purpose), we could outline that according to Guarino (1998) and considering the level of dependence on a particular task, it is possible to distinguish top-level ontologies, domain ontologies, task ontologies and application ontologies.

Table 1: Comparison of lexical resources

CLASSIFICATION CRITERIA	GLOSSARY	DATABASE	DICTIONARY	THESAURUS	LEXICON	ONTOLOGY
Organization	alphabetical order	alphabetical order	alphabetical order	semantically + generically related lexical entries	semantically related lexical entries	semantically related lexical entries
Semantic information	definition in NL	definition + other kinds of info. in NL	definition + pos + etymologies + derivation + usage examples in NL	hierarchical, associative, equivalent relationships	explicit hierarchy (synonymy, antonymy, meronymy...) + grammatical + contextual information	explicitly defined hierarchy relationships around a unique concept
Physical format	paper + electronic format	electronic format	paper + electronic format	paper + electronic format	electronic format	electronic format (readable also by machines)
Domain of knowledge	general + specific	general + specific	general + specific	specific	general + specific	general + specific (agreed by domain experts)

Relevant bibliographic references:

Arano, S.,(2005). “Thesauruses and ontologies” [on-line]. *Hipertext.net*, num. 3, 2005. <http://www.hipertext.net> [Consulted: 15th September, 2006]

Dik, S. C. (1978). *Stepwise lexical decomposition*. Lisse: de Ridder.

Faber, P.B and R. Mairal Usón (1999). *Constructing a Lexicon of English Verbs*. Berlin; New York: Mouton de Gruyter.

Gómez-Pérez, A. Fernández-López, M. and O. Corcho. (2003) *Ontological Engineering*. New York: Springer.

Gruber, T.R. (1993). *Toward principles for the design of ontologies used for knowledge sharing*. [on-line]. Pennsylvania: School of Information Sciences and Technology (IST). Pennsylvania State University. <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html> [Consulted: 21st August 2006].

Martín Mingorance, L. (1998). El modelo lexemático-funcional, [The Functional-Lexematic Model]. Marín Rubiales, A. (ed.). Granada: University of Granada.

Nord (1997). *Translating as a Purposeful Activity*. Manchester: St. Jerome.

Studer, R. et al. (1998). Knowledge engineering: principles and methods. [on-line]. Pennsylvania: School of Information Sciences and Technology (IST). Pennsylvania State University.

Merriam-Webster Online, in <http://www.m-w.com/>

Wikipedia, The Free Encyclopaedia, in http://en.wikipedia.org/wiki/Main_Page

4. Evaluation framework used to compare the lexical resources on each section

This survey is divided into several sections that cover the different kinds of localization approaches of lexical resources and ontologies listed above. The analysis is carried out in a systematic way and all sections are described using the same pattern. Each section includes:

- a) **Short description of the localization approach to be analysed**, which includes information about developers, project duration, and brief description of the approach.
- b) **Comparison of the localization approach against the evaluation framework**. We have designed an evaluation framework which covers the main aspects of the localization process with the aim of obtaining an accurate picture of process and result of the localization strategies used in each approach.

Evaluation framework. The set of criteria used for analysing the following localization approaches is divided into the following items, which will appear in the description of every resource, whenever the corresponding data has been found:

1. **Aims and scope of the localization approach**, which includes information about aims and purpose of the resource, end users, current state of the resource, etc.
2. **Languages and domains involved in the localization process**. In this section we specify the languages in which the resource is already available, and, eventually, the ones into which it will be translated, as well as the domains covered by the resource. In some cases it is even possible to find the number of records in each language, which give us an idea on how consistent is the multilingual resource.
3. **Steps, sources and techniques used for localizing**. This one is the widest section of the evaluation framework, where the translation workflow is detailed, as well as the lexical tools and techniques used for that purpose, when known. The aim is to reel off every step in the localization process conducting to the creation of the consequent multilingual resource.
4. **How multilingual information is displayed**. This section includes screenshots of the resource interface and a description of the search options. The object of this section is to analyse user real options for searching and browsing multilingual data that could eventually be reused in the lay-out of multilingual ontology applications.
5. **Systems of representation of multilingual information**, or how or where multilingual information is stored. Here again –and although it is not an easy aim– the object is to analyse the different options for the storage of multilingual information (Entity-Relation

model diagram or schema), since this problem is currently being tackled in the case of ontologies, and current solutions do not meet all demands

6. **Evaluation methods**, which includes information about the evaluation workflow, when established. The pursue of this section is to report about an eventual automatic or semi-automatic process of the evaluation task, in order to take profit of it for the ontology evaluation task, which is as well a relevant issue in the ontological engineering research.
7. **URL**, where the analysed resource can be accessed.
8. **Contact for information developers**, email addresses of developers.
9. **Relevant bibliographic references**.

5. Glossary localization approaches

5.1 FAOTERM

5.1.1 Short description of FAOTERM

FAOTERM is the multilingual terminology glossary of the Food and Agriculture Organization of the United Nations (FAO) founded in 1945. FAO leads international efforts to defeat hunger. Serving both developed and developing countries, FAO acts a neutral forum where all nations meet as equals to negotiate agreements and debate policy. FAO is also a source of knowledge and information and helps developing countries and countries in transition modernize and improve agriculture, forestry and fisheries practices and ensure good nutrition for all. The FAOTERM system was developed over many years and was first launched on the Internet in January 2001. Nowadays this database consists of approximately 70,000 records.

5.1.2 Comparison of FAOTERM against the evaluation framework

Aims and scope. FAO considered that terminology work is a valuable instrument to support the Organization's role in communications and public information and in creating a common corporate culture.

This role is reflected by the increased demand for terminological tools from all sectors of FAO and its broader 'constituency': (translators, editors, national experts and decision makers, researchers, academics, the media, international organizations, etc.) as well as the will expressed by the governing bodies to strengthen the multilingual capabilities of the Organization.

The increasing amount of multilingual information requires a sound terminology database, not only to provide the correct language equivalents, but especially to standardize terminology within the Organization and within the United Nations system as a whole.

In order to standardize and harmonize the vast quantity of terms used in FAO documents and publications, FAO developed the terminology database FAOTERM and continues to update it with new and current terminology with particular emphasis in emerging areas of work of the Organization such as biotechnology, food standards, avian flu, etc.

Languages and domains. This FAO lexical resource is available in six languages: English, French, Spanish, Arabic and Chinese, including some records in Italian and some which indicate the Scientific Name in Latin. FAO primary role is to provide information on food and agriculture issues, forestry and fisheries.

Out of the total of 70,000 records in FAOTERM, approximately 10,000 records comprise official titles (bodies) of organizations, institutes, programmes, slogans, expert consultations, FAO structure, staff titles...

The records indicate one or more of the 181 main subject areas of FAO's work and searching in FAOTERM may be specified in these domains.

The content of FAOTERM includes all the legacy records over 30 years of collection from FAO documents, publications and glossaries as well as additional content imported following two large Terminology Projects in 2003 and 2005. This meant that over 10 thousand new records in English, French and Spanish and some 50,000 terms were added in Arabic and Chinese. A new Italian collection of approximately 10,000 records were also added to FAOTERM.

Steps, sources and techniques used for localizing. We can identify the following steps in the configuration of the FAO multilingual glossary:

- Manual and automatic screening of FAO documentation and publications, sites, technical reports, etc., and of documents from specific thematic areas of interest to FAO domain of performance, for extracting terminology, which will be then integrated in the MultiTrans' TermBase and Trados WorkBench via Trados MultiTerm (Figure 6), tools which are listed and explained in Table 2. A large part of the entries are created from **English sources**, due to the prevalence of English as source language for document processing in FAO.
- Then, documents are translated, and equivalents to the screened or extracted terms are provided by FAO translators, terminologists, senior revisers, interpreters, editors, technical experts/originators/scientists/collaborators in FAO, as well as counterparts in other organizations outside FAO, especially in the United Nations system (See "External users' consultation" in Figure 6). Localization tools are summarized in Table 2.
- Equivalents can also come from other reliable databases or dictionaries, specially Termium¹¹ and Eurodicautom¹², or other well-known lexical resources (see also "8. International databases" in Table 2).

The whole process of the terminology creation and administration is summarized in Figure 6.

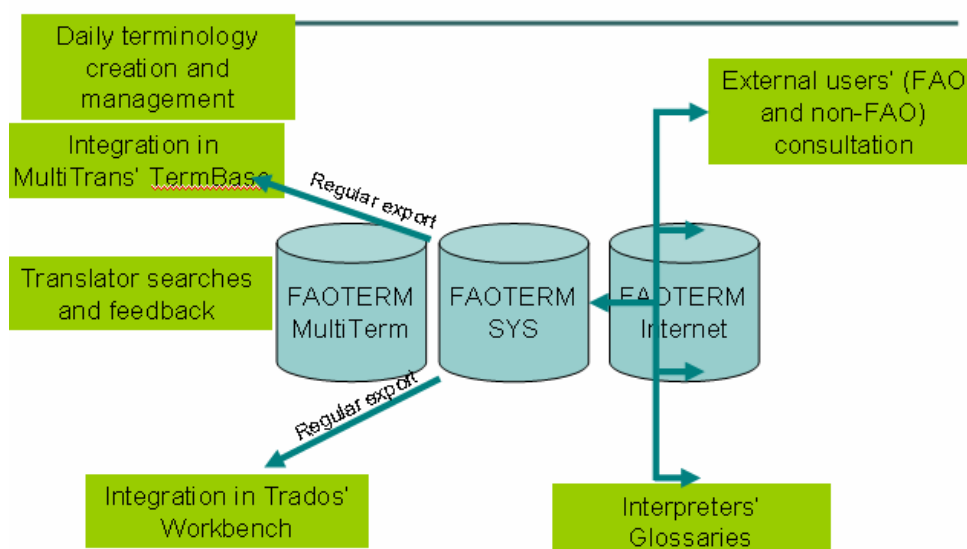


Figure 6: FAOTERM terminology workflow system (TRG/GICM 2006, provided by FAO)

¹¹ <http://www.termium.gc.ca/>

¹² <http://europa.eu.int/eurodicautom/Controller>

Table 2: FAOTERM Localization Tools

Summary of translation tools in FAO	Name
1. Translation memories	Trados Workbench® ¹³
2. Text- and term-bases (aligned multilingual corpora/text-bases for referencing and pre-translation)	MultiTrans of MultiCorpora ¹⁴ (via Trados MultiTerm®)
3. Fulltext search tools	dtSearch® ¹⁵ (used for Arabic correspondence) ISYS v.6 (used for translators to search 1,000 misc. glossaries)
4. Text alignment tools	Trados WinAlign®
5. Term extraction tools	Trados TermExtract® and MultiTrans of MultiCorpora (via Trados MultiTerm®)
6. Glossary building and maintaining tool	FAOTERM System
7. Editor of web pages and other files (including XML files)	Trados TagEditor®
8. International databases	Databases from international organizations as UNTERM ¹⁶ (UN), ILOTERM ¹⁷ (ILO), UNESCOTERM ¹⁸ (UNESCO), SILVATERM ¹⁹ (IUFRO).
9. Feedback systems	E-mail, built in feedback system within FAOTERM itself.
10. Fora/Networks	JIAMCATT (Joint Inter-Agency Meeting on Computer-Assisted Translation and Terminology, a restricted inter-agency forum of approximately 80 institutions for the exchange of glossaries, files, ideas and discussion of terminology, translation and their management).
11. Internet resources	The vast array of linguistic resources available on the Internet.

Output Strategy/Subset delivery

An advanced “**administrators’ module**” allows for batch exports and imports in various formats (Word, Excel, HTML, XML, MultiTerm) and automatic generation of **multilingual glossaries** using special filtering options. Administrators are able to prepare project-related, subject-related, or meeting-related glossaries at the click of a button. The module also includes full **forecasting and reporting features** for a more automated approach to terminology management. **Security features** such as batch publishing and backup of data have been introduced with timing

¹³ <http://www.trados.com/products.asp?page=1214>

¹⁴ http://www.multicorpora.ca/index_e.html

¹⁵ <http://www.dtsearch.com/index.html>

¹⁶ <http://unterm.un.org/>

¹⁷ <http://www.ilo.org/iloterm/>

¹⁸ <http://termweb.unesco.org/>

¹⁹ <http://193.170.148.70/silvavoc/search.asp>

mechanisms.

Also, a “**What’s New**” feature allows users to review the latest titles and terms which have been introduced.

Multilingual information display

As shown in Figure 7, FAOTERM offers us the possibility of looking for a term in one or more of the six source languages. Results can also be shown in all languages or just in the selected ones. The query can be a word, an expression, an abbreviation, etc. Category (bodies or terminology) and subject can as well be determined. After doing the search a hit list of matching terms will be displayed on the left of the screen. From the hit list, single entries can be clicked, which will then be shown on the right with the following information:

- Record information: entry number; category, status, reliability, source language (usually English), source (full bibliographic details, codes, URLs, etc.) and subject.
- Linguistic information: results in the selected languages and term source.

The screenshot displays the FAOTERM web interface. At the top, the FAO logo and the text 'FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS' are visible, along with the tagline 'helping to build a world without hunger'. Below this is the 'FAO TERMINOLOGY' header and a navigation menu with links for Home, Site Map, Contacts, and AGROVOC. A search bar is present with the text 'saline water' entered. To the right of the search bar are buttons for 'Search' and 'Clear', and a 'Category:' dropdown menu set to 'ALL'. Below the search bar are several checkboxes for search options: 'Exact match', 'Fuzzy search', 'Match any word', and 'Also search in definition'. A 'Search in AGROVOC' button is also visible. Underneath, there is a 'Target languages:' section with checkboxes for 'ALL', 'العربية', 'English', 'Español', 'Français', 'Italiano', and 'Scientific name'. The search results are displayed in a list on the left, showing 'Found 2 item(s)'. The first result is 'saline water conversion process [IRRIGATION]' and the second is 'saline water reclamation [ENVIRONMENT]'. On the right, a detailed view of the search results is shown, including the term 'saline water reclamation' in English, its Spanish equivalent '(tratamiento de) desalinización de agua (salada)', and its French equivalent 'dessalement de l'eau'. The Arabic equivalent 'تخية المياه المالحة' is also shown. The source information for each term is provided, such as 'International Centre for Agricultural Research in the Dry Areas, ICARDA, FAO Terminology Project, 2003.' for the English term.

Figure 7: FAOTERM interface

Systems of representation of multilingual information

FAO TERM is based on the XML²⁰ Its data representation is presented in the following format (see below). This format has been generated from a database schema, and a part from administrative data as the “creation date” or the “author” of the information input, what we can observe is that there exists a “term” that belongs to a “category” and to a “subject”, and that has been recovered from a “source”, and that is defined in the “definition” or “gloss” category. Following this information we found the different terms or names for each language of the resource (identified by the entity “term”) accompanied by the attribute langSet xml:lang that acquires a different value depending on the language in question (e.g. <langSet xml:lang=”ar”> for the Arabic language). Because of the form in which multilingual information is stored, i.e. following well-known standards, its reuse and exchange with other resources is guaranteed. Such standards will have to be considered for the representation of multilingual information in ontologies

```
<termEntry id="tid-FAO-45827">
  <wf-step>4</wf-step>
  <transacGrp>
    <transac type="origination">super</transac>
    <date>1999-04-13T18:03:46</date>
  </transacGrp>
  <transacGrp>
    <transac type="modification">alldritt</transac>
    <date>2006-05-31T14:53:25</date>
  </transacGrp>
  <descrip type="Remarks" order="10">Previous title: World Wildlife Fund</descrip>
  <descrip type="Category" order="1">Bodies</descrip>
  <descrip type="Subject" order="9">TITLES</descrip>
  <descrip type="Source" order="8">Yearbook of Int. Org., 1988/89; GLOSSFORBOD</descrip>
  <descrip type="Headquarters" order="7">Gland, Switzerland</descrip>
  <descrip type="Glossint" order="11">Commission on Genetic Resources for Food and
Agriculture</descrip>
  <langSet xml:lang="ar">
    <tig>
      <term>الصندوق العالمي لحماية الطبيعة</term>
    </tig>
  </langSet>
  <langSet xml:lang="en">
    <tig>
      <term>World Wide Fund for Nature</term>
    </tig>
  </langSet>
  <langSet xml:lang="">
    <tig>
      <term></term>
    </tig>
  </langSet>
  <langSet xml:lang="es">
    <tig>
      <term>WWF</term>
      <descripGrp>
        <descrip type="Form">Abbreviation</descrip>
      </descripGrp>
    </tig>
    <tig>
      <term>Fondo Mundial para la Naturaleza</term>
    </tig>
  </langSet>
  <langSet xml:lang="fr">
    <tig>
      <term>Fonds mondial pour la nature</term>
    </tig>
    <tig>
      <term>WWF</term>
      <descripGrp>
        <descrip type="Form">Abbreviation</descrip>
      </descripGrp>
    </tig>
  </langSet>
  <langSet xml:lang="zh">
    <tig>
```

²⁰ <http://www.w3.org/XML/>

```
<term>世界大自然基金</term>
</tig>
</langSet>
</termEntry>
```

Evaluation methods. The System has been redesigned to include a fully **flexible workflow system** with profiles for editors, validators, post-validators and publishers, providing a formal controlled input with linguistic control. The System has been designed to include direct contributions from **collaborative partners** and **Web Services** are envisaged in a future phase.

URL: <http://www.fao.org/faoterm/map.asp?lang=EN&open2=1&what=1>

Contact for information developers: http://www.fao.org/UNFAO/about/index_en.html

Relevant bibliographic references:

<http://www.fao.org/faoterm/index.asp?lang=en>

6. Database localization approaches

6.1 FishBase

6.1.1 Short description of the FishBase

Antecedents of the FishBase database and glossary were the FAO publications *Identification Sheets* (Fischer 1973) and *FAO Species Synopses* and *FAO Species Catalogues* (Fischer 1976). These works inspired experts throughout the world to elaborate and collaborate on the production of fish species catalogues and databases. In 1994 a global database of basic information on fish and invertebrates, the SPECIESDAB, was also developed by FAO (Coppola *et al.* 1994). It was then when FishBase was conceived by Daniel Pauly in 1987. Pauly intended to create a database which would be continuously updated and available to others in what was then known as the 'ICLARM Software Project (Pauly *et al.* 1995). Rainer Froese incorporated to the project in 1988 and suggested the implementation of the database in DataEase, which would be the birth of the current database.

The original database and glossary were available in English, but the necessity to communicate and to make information in FishBase available to people around the world, led to an early initiative to provide translations of FishBase in the major languages used in Africa, the Caribbean and the Pacific.

6.1.2 Comparison of FishBase against the evaluation framework

Aims and scope of the localization approach. The FishBase multilingual database was developed in order to unify the terminology in the field of ichthyology and fisheries, and become in this way a reliable source of information and communication between experts all over the world. FishBase includes 29,400 species, 222,300 common names, 42,600 pictures of fishes and 38,600 references.

Languages and domains involved in the localization process. Fish species can be looked for in the following languages: English, Spanish, Portuguese, French, German, Italian, Dutch, Chinese, Italian, Greek, Swedish, Russian, Farsi Vietnamese, Thai, Bahasa Malay/Indonesian. The species name in Latin appears always next to the different names given to the fish sort in the different regions of the world, as well as the region of origin of the species. As shown in Figure 8,

free-text fields -information fields which are not always included but just when the corresponding information is available, as for example, those reporting on the size, environment, climate or biology of the species- were originally compiled in English and subsequently translated into the rest of languages. However, when language versions different from English are requested, the English version still appears (cf. Figure 9).

Domains involved in the database are ichthyology and fisheries.

Steps, sources and techniques used for localizing. In order to tackle the enormous task of translating the English database to the many different languages, a strategy was proposed by the FishBase team consisting on three different phases:

- 1) Translation of **terms** and **definitions** from English into French, Portuguese and Spanish, to start with, and then to the rest of the database languages. This task was carried out by collaborators of the FishBase team, who were often not linguists but native speaker experts in the field. LRs that supported the localization process were mainly:
 - o **GLOSSARY** table: available in annual versions of the FishBase CD-ROM from 1996 to the present.
 - o Translations of the *FishBase 98 book* (<http://filaman.ifm-geomar.de/contents.htm>) from English into French, Portuguese, Spanish.
Both resources were as well translated by native speaker experts in the field working as collaborators in the FishBase team (<http://www.fishbase.org/FBTeam.cfm>).
- 2) Translation of **fixed text** (title, labels, notes) of the web page to the resource languages. In Figure 8 fixed text is to be found in bold characters on the left column.
- 3) Simplification and standardization of vocabulary and grammar in English **free-text** fields (text on the right in Figure 8) to achieve good results with the use of the machine translation (MT) systems of the European Commission, ECMT (European Commission's Machine Translations Service) and Systran²¹. Possible language combinations for this MT system are English into Dutch, French, German, Greek, Italian, Portuguese and Spanish. Missing topics in the dictionary library of the MT systems were solved by the implementation of special dictionaries or glossaries (the FishBase GLOSSARY and the FishBase successive books) compiled by FishBase and added to the Systran dictionary.

How multilingual information is displayed. After having introduced the search term in English (a fish name by its common name or scientific name in Latin) in the main page, a list of the common names attributed to the species is displayed. By clicking on the species name, FishBase opens the **Species table**, where information as family name, order, class, size, climate, biology, etc. is displayed, as shown in the figure below (Figure 8).

²¹ <http://www.systransoft.com/index.html>

Poecilia gillii [\(Kner, 1863\)](#)

Family: [Poeciliidae](#) (Poeciliids), subfamily: Poeciliinae

Order: [Cyprinodontiformes](#) (rivulines, killifishes and live bearers)

Class: [Actinopterygii](#) (ray-finned fishes)

FishBase name:

Max. size: 6.0 cm TL (male/unsexed; Ref. 36880); 10.5 cm TL (female)

Environment: benthopelagic; pH range: 7.2 – 7.8; dH range: 10 - 30

Climate: tropical; 24 – 28°C

Importance:

Resilience: Medium, minimum population doubling time 1.4 - 4.4 years (Preliminary K or Fecundity.)

Distribution: Central and South America: Atlantic drainage from Guatemala to the Río Térraba, Costa Rica. Río Grande, Coclé Province to the Río Bayano, Panama.

[Gazetteer](#)

Biology: Found in waters of all current velocities, but are more abundant in slack waters. Very large individuals (up to 105 cm) are found in brackish water, while smaller individuals occur in brooks to an elevation of 1220 m. Inhabit swamps, brooks and in shallow waters of large rivers, generally found near the substratum browsing on detritus, ooze and filamentous algae, reproduces throughout the year, but with a peak in juvenile abundance in August (Ref. 36880).

Red List Status: (Ref. 53964)

Dangerous: harmless

Coordinator:

Main Ref: [Rodriguez, C.M., 1997. \(Ref. 26855\)](#)

[Update](#) | [Add](#) | [Get XML file](#) | [Point data in XML](#) | [Common names in XML](#) | [Photos in XML](#)

More information:

Countries	Common names	References	Collaborators
FAO areas	Synonyms	Growth	Genetics
Occurrences	Pictures	*L-W relationship	Allele frequencies
Introductions	Sounds	L-L relationship	Heritability
Ecosystems	Reproduction	Length frequencies	Strains



[Point map](#)

Figure 8: Search result for the *Poecilia gillii* species known commonly as “molly” in Costa Rica

As already explained in **Languages and domains involved in the localization process** section, by consulting the database in any other of the resource languages except for English, **terms, definitions and fixed text** are presented in the selected language. However, **free text** sections are displayed in English -as can be observed in Figure 9 for the Spanish version of the searched species *Rainbow trout*- accompanied by the corresponding translation coloured in blue. In addition, a translation section is included which specifies the European Commission machine translation service used for the translation, the ECMT in the mentioned example.

Translation:	The following three fields were machine-translated by ECMT . You can also use SysTran .
Distribución: Gazetteer	<p>Eastern Pacific: Kamchatkan Peninsula and have been recorded from the Commander Islands east of Kamchatka and sporadically in the Sea of Okhotsk as far south as the mouth of the Amur River along the mainland. The records outside of Kamchatka probably represent migrating or straying Kamchatkan steelhead (<i>penshinensis</i>) rather than the established native population (Reg. 50080). One of the most widely introduced fishes, may be regarded as global in its present distribution. In the tropics restricted to areas above 1,200 m. Several countries report adverse ecological impact after introduction (Ref. 5723).</p> <p>Pacífico del este: Se ha registrado una península del Kamchatka y del este de las islas del comandante de Kamchatka y esporádicamente en el mar de Ojotsk como el Sur Lejano como la boca del río Amur a lo largo del continente. Los documentos fuera de Kamchatka representan probablemente a trucha arco iris del Kamchatka de migración o de perdición (<i>penshinensis</i>) en vez de la población nativa establecida (Reg. 50080). Uno de los peces lo más ampliamente posible introducidos, puede considerarse como global en su actual distribución. En los trópicos restringidos a áreas sobre 1.200 m. Varios países informan el impacto ecológico adverso después de la introducción (ref. 5723).</p>
Morfología:	<p><u>Dorsal spines</u> (total): 3 - 4; <u>Dorsal soft rays</u> (total): 10 - 12; <u>Anal spines</u>: 3 - 4; <u>Anal soft rays</u>: 8 - 12; <u>Vertebrae</u>: 60 - 66. Body elongate, somewhat compressed especially in larger fish. No nuptial tubercles but minor changes to head, mouth and color occur especially in spawning males. Coloration varies with habitat, size, and sexual condition. Stream residents and spawners darker, colors more intense. Lake residents lighter, brighter, and more silvery. Caudal fin with 19 rays (Ref. 2196).</p> <p>♂: 3-4; ♀: 10-12; 3-4; : 8-12; : 60-66. cuerpo elongate, un tanto comprimido especialmente en peces más grandes. Ningún tubérculo nupcial pero los cambios de poca importancia a la cabeza, a la boca y al color ocurre especialmente en varones de freza. La coloración varía con el hábitat, el tamaño, y la condición sexual. Residentes y spawners de corriente más oscuros, colores más intensos. Residentes de lago más ligeros, más brillantes, y más plateados. Aleta caudal con 19 rayos (ref. 2196).</p>

Figure 9: Fragment of the Species table for *Rainbow trout* in Spanish

Although it has already been said that the FishBase languages are English, Spanish, Portuguese, French, German, Italian, Dutch, Chinese, Italian, Greek, Swedish, Russian, Farsi Vietnamese, Thai, Bahasa Malay/Indonesian, translations of the **free text** sections are only available for those languages supported by the European Commission MT services.

Another option is to check a term in the **GLOSSARY** in English, French, Spanish, Portuguese and Russian. The glossary additionally offers the definition of the term and links to related terms or other related on-line glossaries (see Figure 10).

Term : pálpebra			
<i>Language</i>	<i>Term</i>	<i>Definition</i>	<i>See also</i>
English	eyelid	Moveable, muscular fold of skin capable of covering all or part of the exposed portion of the eyeball.	eye
French	paupière	Pli mobile, musclé de peau capable de couvrir tout ou une partie de la portion exposé du globe de l'oeil.	oeil
Spanish	párpado	Pliegue muscular móvil de la piel, capaz de cubrir total o parcialmente la porción expuesta del globo ocular.	ojo Web de la búsqueda
Portuguese	pálpebra	Prega cutânea muscular e móvel que cobre toda a área exposta do globo ocular.	olho
Russian			
Cyrillic	веко		вытапливание. вытеснение

Search this term in other glossaries.
Use the 'Return' button of your browser to return to FishBase.

EEA Glossary:	Glossary of the European Environmental Agency.
Encyclopedia Britannica:	Highly authoritative source.
EPA Terms of Environment:	Glossary of the U.S. Environmental Protection Agency.
Tree of Life:	Glossary of biological terms.
LingInfo:	Glossary of molecular biology terms.
Google Search:	Try the largest Internet Search Engine.
Google Images:	Search for Images related to the Term.
Dictionnaire Universel Francophone:	In French.
Grande Dicionario Universal Lingua Portuguesa:	In Portuguese.

Search new term:

Figure 10: FishBase GLOSSARY

Systems of representation of multilingual information. No information has been found referring to this, in spite of contacts established to developers. However, it can be assumed that information is stored in a database, as the resource name suggests.

URL: <http://www.fishbase.org/search.php?lang=English>

<http://www.fishbase.net/>

Contact for information developers. To have access to e-mail addresses and/or telephone numbers from developers of FishBase and members or former members of the team, consult the following web address.

<http://www.fishbase.org/FBTeam.cfm>

Relevant bibliographic references

<http://www.fishbase.org/manual/English/contents.htm>

7. Dictionary localization approaches

7.1 Eurodicautom

7.1.1 Short description of Eurodicautom

Eurodicautom is the multilingual online dictionary of the European Commission (EC)²². It was first set up in 1973 and it was the result of the cooperation work of terminologist, translators and computer science experts of the European Commission. Eurodicautom was originally developed mainly to solve the needs of the translators working for the European institutions (see Directorate General for Translation - DGT²³). Rapidly it became a very useful tool and was adopted by linguists in other European institutions. Nowadays it is a free available LR on the Internet and receives an average of 120,000 enquires every day. Terminologist and linguists of the DGT are constantly updating it. At present the term bank contains about five and a half million entries (terms and abbreviations), subdivided into more than 800 collections.

7.1.2 Comparison of Eurodicautom against the evaluation framework

Aims and scope. Eurodicautom meets the demands of the European Union (EU) objective of giving every official language the same recognition. That is why terminology in the EU is so important, and more particularly within the EC, since this is the organism responsible for EU citizens obtaining the adequate information about the EU policy in their own language. To meet those terminological needs, the Terminology Unit has a team of terminologists who are in charge of enlarging and updating the Commission's large dictionary, Eurodicautom, in order to help translators to solve their terminology problems.

Languages and domains. Eurodicautom covers twelve languages, eleven official languages (Danish, Finnish, Greek, Portuguese, Dutch, French, Italian, Spanish, English, German and Swedish) and Latin, containing five million terms and two hundred thousand abbreviations. All languages are not equally represented: those languages of the founder countries have more entries than the more recently-added languages. Consultations can be carried out from any source language into one or more target languages.

Eurodicautom includes lexical entries related to many domains of the human knowledge, but it is particularly rich in technical and specialised terminology related to EU policy (agriculture, telecommunications, transport, legislation, finance). Entries are classified into 48 subject fields, as for example, medicine or public administration, and each of them constitutes a technical dictionary.

Steps, sources and techniques used for localizing.

- The first steps taken for the elaboration of a common Dictionary for translators of the EC were carried out by the translators themselves, as they used to elaborate technical cards of every technical term they came across. **Source languages** were mainly French and English, since these are the languages in which the EU documents are first drawn up.
- Afterwards, two lexical tools were merged to become the foundations of the Dictionary, DICAUTOM -a phrasal automatic dictionary launched in 1962 in the four languages official at that time (French, German, Italian and Dutch) - and EUROTERM published in 1964 -a phraseology dictionary available in the same four languages.
- The Terminology Bank of the University of Montréal, Canada, put 80,000 bilingual cards (English-French) at the disposal of the EC.

²² <http://ec.europa.eu/>

²³ http://ec.europa.eu/dgs/translation/index_en.htm

- Other glossaries were as well merged (Goffin 1997), as well as resources from other European and national institutions, which were used for enriching what would become the final document.
- In 1976 Eurodicautom was finally launched as a multilingual automatic dictionary, and when more countries joined the EU, the dictionary had to be enlarged continuously by a team of terminologists, specialized in the task.
 - The enlargement was made mainly **manually** by terminologists. Multilingual information was extracted from the multiple publications of the EC, especially from the Official Journals (manually at the beginning, **semi-automatic** in the recent years). This work was supervised by experts in the corresponding domain. Translators also contributed to the task by delivering to the Terminology Unit computerized terminological cards developed for this purpose, whenever new terms were introduced and translated within the EU institutions.
 - In 1995, with the introduction of the **Euramis**²⁴ project (*European Advanced Multilingual Information System*), a series of e-mail based, client-server applications, became automatic and enabled translators to a more effective management of terminology, which would be used for widening Eurodicautom. These applications provided access to a variety of services in the field of natural language processing - **translation memories** (Trados Multi Term²⁵), **mass processing of linguistic data**, **machine translation** (Systran²⁶), and **workflow automation**- the store and management of term bases.

Multilingual information display. In this section we describe the user interface of Eurodicautom. In the initial search, the user can select the source language, the subject, the target language or languages and the way in which the information is to be displayed, as can be seen in Figure 11.

²⁴ See for a more detailed information
http://ec.europa.eu/translation/reading/articles/pdf/1998_01_tt_blatt2.pdf#search=%22euramis%22

²⁵ <http://www.trados.com/>

²⁶ <http://www.systransoft.com/index.html>

IMPORTANT LEGAL NOTICE: The information on this site is subject to a disclaimer and a copyright notice.

Welcome to Eurodicautom, the multilingual term bank of the European Commission.

Please note that, because of the migration to the new interinstitutional database IATE, Eurodicautom will no longer be updated. We shall keep you informed about further developments.

Figure 11: Eurodicautom interface

To obtain the whole information available in the dictionary we choose **All fields** in the **Display** section on the first page. Those results to a search in which all fields should be displayed offer the user the following information:

- **Hit list**, in which the searched term appears -alone or being part of a compound-, as well as other semantically related terms. In this section we also find internal information about the terminology office that has introduced the terminology data about a term (for example BTL which is equivalent to “Terminology Office, European Commission Luxemburg”) and the identification number.
- **Document** section. In this section of the dictionary (to be seen in Figure 12) we find the searched **term** -in the language of the search- and the possible translation in the selected languages. In this section also the **Subject** and the **Reference** sub-sections accompany the searched term and the corresponding translations. The **Reference** gives us a hint about the reliability of the results; since we can check if it is an authoritative source. The **Subject** sub-section offers information about the specific subject field of the term, which is marked with an abbreviation that represents the general domain out of a total of 48 in which the dictionary is divided.

In case we should not need information about the subject, the reference or the definition, we could choose the mode **Terms** in the **Display** section of the first page.

[User Guide](#)

HitList	Extend	New Query
1.	contamination(1) pollution(2)	BTL - CNB69 - 820
2.	pollution(1)	BTB - CIF90 - 1084
3.	contamination(1) pollution(2)	BTL - AIR94 - 1000086

Document 1		Extend	New Query	Feedback
<i>Subject</i> Nuclear Technology - Nuclear Industry (AT) Defence - Warfare (DE)				
EN				
(1)	TERM	contamination		
	<i>Reference</i>	AFNOR;Dict.techn.ill.IV-60		
(2)	TERM	pollution		
	<i>Reference</i>	AFNOR		
DE				
(1)	TERM	Verunreinigung		
	<i>Reference</i>	Dict. Techn. ill. IV-60		
(2)	TERM	Verseuchung		
	<i>Reference</i>	Dict. Techn. ill. IV-60		
ES				
(1)	TERM	contaminación		
	<i>Reference</i>	AIPCN, Dicc técnico ilustrado, IV		

Figure 12: Results for a searched term in Eurodicautom

Systems of representation of multilingual information. No information has been found referring to this, in spite of contact established to developers.

Evaluation methods. The evaluation is carried out manually by terminologist in the Terminology Unit and translators, who are the end users of the dictionary. Feedback from users outside the European institutions is welcomed and can be done from the “Document” section page (see Figure 12) by clicking on “Feedback” and sending an email to the Terminology team.

URL: <http://ec.europa.eu/eurodicautom/Controller>

Contact information for developers: DGT-EURODICAUTOM-INT@cec.eu.int

NOTE: Since January 2007, the above mentioned URL refers to the following one: <http://iate.europa.eu/iatediff/>

Eurodicautom has been imported into the IATE (InterActive Terminology for Europe), inter-institutional terminology database system, which merges all EU terminology resources. Despite this replacement, the analysis of the localizing process of Eurodicautom is still valid for the purposes of this survey.

Relevant bibliographic references:

Goffin, R. (1997): « EURODICAUTOM. La banque de données terminologiques multilingues de la Commission européenne (1973-1997) » in *Terminologie et Traduction* 2.1997, 30-73.

González, L and P. Hernández: La terminología en la Comisión Europea Link: <http://www.termilat.info/public/env100.doc> [Accessed in September 2006]

Hernández, P. (2000) Las bases de datos terminológicos de la Comisión Europea. EURODICAUTOM. En Gonzalo García, C. y García Yebra, V. eds: 2000: *Documentación, Terminología y Traducción*. Madrid: Síntesis, Fundación Duques de Soria: 97-107.

Directorate-General for Translation of the European Commission, April 2005, *Translating for a Multilingual Community*, in http://ec.europa.eu/dgs/translation/index_en.htm

Directorate-General for Translation of the European Commission, July 2005, *Translation Tools and Workflow*, in http://ec.europa.eu/dgs/translation/index_en.htm

8. Thesauri localization approaches

8.1 AGROVOC

8.1.1 Short description of AGROVOC

The AGROVOC Thesaurus was developed by the Food and Agriculture Organization (FAO) and the Commission of the European Communities in the early 1980s, and first published in 1982 in three languages: English, Spanish and French. It is defined as a multilingual structured and controlled vocabulary. In the following URL http://www.fao.org/aims/ag_figures.jsp the number of terms per language is calculated real time. At the moment of the query (26/09/2006) Spanish was the language with the highest number of terms with 41,580 terms.

8.1.2 Comparison of AGROVOC against the evaluation framework

Aims and scope. AGROVOC's aim is to standardize the indexing process in the agricultural domain in order to make searching simple and more efficient, and to provide the user with the most relevant resources. AGROVOC is currently used for indexing and retrieving data in agricultural information systems inside the FAO (e.g. the international information system for the agricultural sciences and technology, AGRIS/CARIS²⁷) and outside this organization.

Languages and domains. AGROVOC was first created by domain experts in agriculture in English and then manually translated to Spanish and French. Nowadays, it is available online in 10 languages (English, French, Spanish, Arabic, Chinese, Japanese, Portuguese, Thai, Czech, Slovak). It will be soon also online available in Thai, Lao and Hindi. It is in development for Marati and other 2 Indian languages, Polish, Korean, Farsi, Hungarian and Malay. It is also under revision for Italian and German, and the Amharic, Catalan and Russian communities have expressed interest for a translation.

AGROVOC is used for the description of sources in the field of agriculture, forestry, fisheries, nutrition, food and related domains, e.g. environment and sustained development, among others.

Steps, sources and techniques used for localizing. The fact that the different language versions of AGROVOC are generally carried out by translating the English version means that the localization approach is centred on the semantic of the English words. Each version of the AGROVOC thesaurus is carried out by native speakers (terminologists and translators) in the corresponding country, and that is why the translation workflow cannot be exactly defined. In general terms we identify the following steps in the localizing task:

²⁷ <http://www.fao.org/AGRIS/>

Step 1: Translators or terminologists have access at a specific FAO resource called FAOTERM²⁸, which is a glossary with translations in 5 languages, as the first resource in the search task for the aimed translation.

Step 2: Research in agricultural resources in order to see the current and real use of a term (e.g. AGRIS resources).

Step 3: Search on a list of existing online LRs made available by FAO which include (see Appendix 1 for the complete detailed list):

- Multilingual Thesauri (as for example: UNESCO, UNBIS Thesaurus, CAB Thesaurus)
- Lexicons (e.g. WordNet)
- Dictionaries (e.g. The Dictionary of Agricultural Occupations)
- Encyclopaedias (Wikipedia)

Step 4: Consultation of guidelines for thesauri development and translation established by FAO:

- Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies, ANSI/ISO Z39.19-2005;
- Guidelines for the establishment and development of multilingual thesauri, ANSI/ISO 5964-1985.
- Guidelines for Multilingual Thesauri, Working Group on Guidelines for Multilingual Thesauri, Classification and Indexing Section, IFLA, April 2005;
- The

FAO	House
-----	-------

 style (<http://www.fao.org/docrep/004/AC339e/AC339E00.htm>)

Multilingual information display. In order to browse a term in the AGROVOC thesaurus (Figure 13), a word block (made up of one or more words) in a specific language is introduced, “Fish” in the screenshot,. All block words containing that word are displayed. By selecting one of the results, “Fish” again, translations of the word in the rest of languages are displayed (Figure 14), just as the hierarchical and non-hierarchical relations to other terms in the original language of the search: **BT** (broader term), **NT** (narrower term), **RT** (related term), **UF** (non-descriptor). **Scope notes** appear at the end of the term list and are used to clarify meaning or context of terms (In Figure 14, the explanation “Use for fish as food; for the animal use *Fishes*”, refers us to “Fishes” if we want to find equivalents for another meaning of “Fish”).

²⁸ www.fao.org/faoterm, also analyzed in this document under **Glossary localization approaches**, 4.1 FAOTERM

- **AGROVOC Thesaurus**
 - Browse
 - Sub-vocabularies
 - Latest updates
 - Suggest terms
 - Download
 - Webservices
 - Copyright information
- **Knowledge Organization Systems**
 - By Type
 - By Subject area
- **AOS/CS**
 - The Concept Server
 - Applied ontologies in FAO
 - Ontology relationships
- **Glossary**
- **Frequently Asked Questions**

AGROVOC Thesaurus Last Update: 11/07/2006

AGROVOC is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment).

Search term:

starting with
 containing text
 exact match

Search results for terms containing: fish

Terms found: 80 Pages: [1](#) [2](#) [3](#) [4](#) [Next >>](#) [Last](#)

Fish	15903	EN	Descriptor with relations
Fish air bladder	15925	EN	Non-Descriptor with USE relation
Fish cages	29786	EN	Descriptor with relations
Fish catching	10875	EN	Non-Descriptor with USE relation
Fish conversion	32447	EN	Non-Descriptor with USE relation
Fish culture	2918	EN	Descriptor with relations
Fish detection	2919	EN	Descriptor with relations
Fish diseases	2920	EN	Descriptor with relations
Fish extracts	2921	EN	Descriptor with relations
Fish factories	28608	EN	Descriptor with relations

Terms in this page: 10

Figure 13: Interface of the AGROVOC Thesaurus, 1st step in the search

- Webservices
- Copyright information
- **Knowledge Organization Systems**
 - By Type
 - By Subject area
- **AOS/CS**
 - The Concept Server
 - Applied ontologies in FAO
 - Ontology relationships
- **Glossary**
- **Frequently Asked Questions**

Search term:

starting with
 containing text
 exact match

EN : Fish	BT : Animal products
FR : Poisson (aliment)	BT : Fishery products
ES : Pescado	RT : Perishable products
AR : أسماك (غذاء)	RT : Foods
ZH : 鱼	RT : Seafoods
PT : Carne de peixe	RT : Fresh products
CS : rybí maso	RT : Postmortem changes
JA : 魚	RT : Fish products
TH : ปลา (อาหาร)	SNR : Fishes
SK : ryba (mäso)	SNX : Fishes
	UF : Fresh fish
	UF : Wet fish
	UF : Fish meat

Scope Note : Use for fish as food; for the animal use "Fishes" (2943)
Term code: 15903

Figure 14: Interface of the AGROVOC Thesaurus, 2nd step in the search

Systems of representation of multilingual information . The figure below (Figure 15) pictures the tables and fields of the AGROVOC thesaurus in a relational database format.

The main tables of the database and a short description of the relevant fields for the purpose of this work are summarized as follows:

1. **agrovocterm**: This table contains all AGROVOC terms and the code assigned to them, which will be the same for all realizations in the different languages. The "termcode" field provides the link to the **termlink** table. The "languagecode" field contains the information of the code assigned to the term being described and is the link to the **language** table. Finally, the "termspell" field supplies the lexicalization of the term in the specific language.
2. **termlink**: This table contains all relationships among the terms.
3. **termtag**: This table is used to generate the scope notes for each language for terms.
4. **tagtype**: This table is a reference table for the Tag type.

5. **scope**, **termstatus** and **linktype**: Reference tables for scope, status of terms and types of relationships respectively.
6. **language**: Reference table for languages used in AGROVOC.
7. **mapping**: This table maps AGROVOC terms to the AGRIS/CARIS categories.
8. **maintenancegroup**: Table containing information about the owners of the AGROVOC terms.
9. **catschemes**: Table containing information about additional classifications schemes.
10. **categories**: Table containing information about categories.

Each represented term has associated a term code and a language code. One and the same term code is shared by all “equivalent” terms in the different languages. The language code refers to a separate label where all available languages are listed.

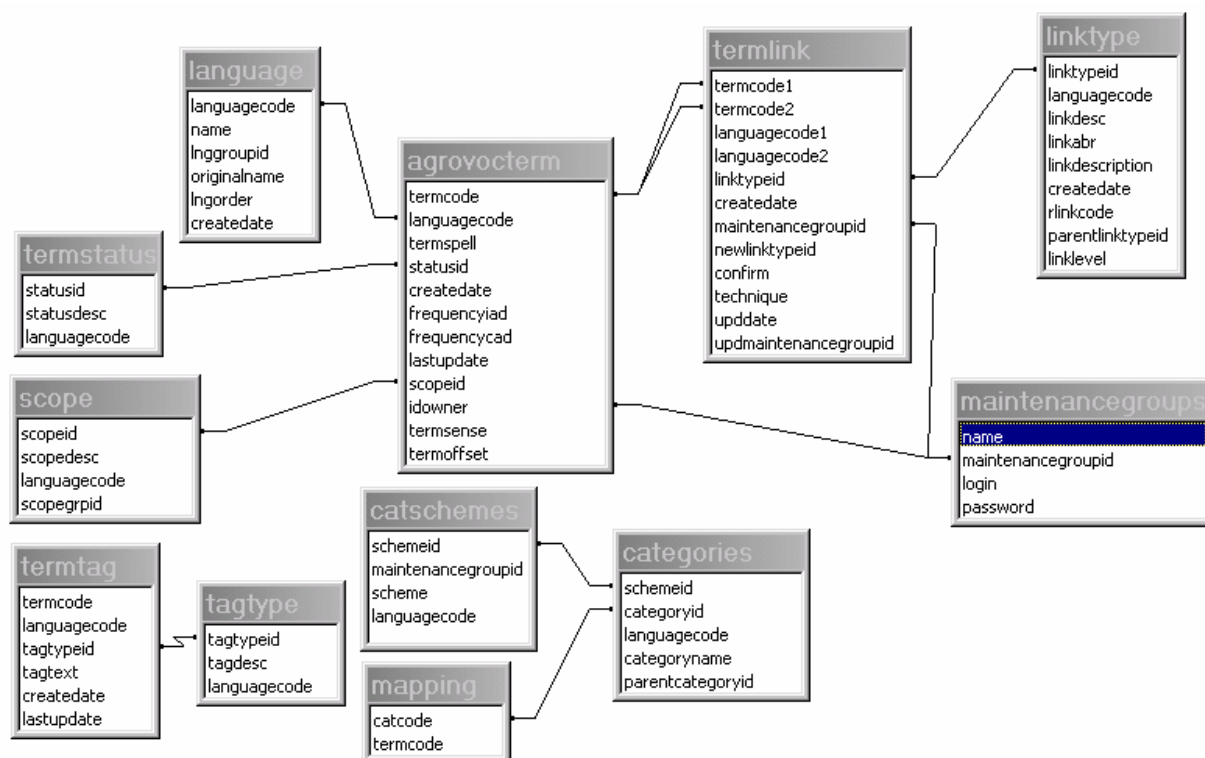


Figure 15: AGROVOC Systems of representation of multilingual information

Evaluation methods. Evaluation is done semi-automatic, making use of two tools developed by FAO for the maintenance and refinement of AGROVOC.

- **AGROVOC Maintenances Interface** is a system that allows for interaction with the database storing the AGROVOC thesaurus, and although the thesaurus is multilingual, the tool interface is currently available only in English (technical details: PHP, MySQL).
- **AGROVOC Thesaurus refinement tool** presents an approach to detect ill-defined relationships of terms and suggest more precisely ones. The system consists of three main modules: Refinement rule Acquisition, Detection and Suggestion, and Verification. The rule acquisition module is a tool used for acquiring the refinement rules from experts and by machine learning. The Detection and Suggestion module uses noun phrase analysis with WordNet alignment to detect inappropriate relationships and to make suggestions for more

appropriate ones based on the rules of acquisition. The Verification module is a tool for verifying and for confirming.

FAO has been working for some time now on **restructuring** the AGROVOC thesaurus which envisages a preliminary revision and multilingual enrichment phase, followed by a semantic restructuring phase (see <http://www.fao.org/agris/aos>). A restructuring of the thesaurus is planned in order to abandon the term-oriented approach in favour of a clear distinction between concepts, terms and strings, and a clear, and most important, distinction between cultural perspectives, not reflected by the English oriented view (see Appendix 2).

URL: <http://www.fao.org/aims/>

Accessed September 2006

Contact for information developers: FAO-Agris-Caris@fao.org.

Relevant bibliographic references:

AGROVOC Thesaurus maintenance and refinement tools: http://www.fao.org/aims/tools_thes.jsp

FAO's Role in Information Management and Dissemination-Challenges, Innovation, Success, Lessons Learned: <http://www.fao.org/docrep/008/af238e/af238e04.htm>

Reengineering thesaurus for new applications: AGROVOC example, article in:
<http://journals.tdl.org/jodi/article/viewArticle/jodi-126/111>

8.2 Eurovoc

8.2.1 Short description of Eurovoc

The European Communities and the Office for Official Publications of the European Communities started working on Eurovoc at the end of the 70's and the first version in 7 languages was first published in 1984. Eurovoc is currently being used mainly by the European Parliament, the Office for Official Publications of the European Communities, and the national and regional parliaments in Europe. This thesaurus contains currently about 70,000 descriptors.

8.2.2 Comparison of Eurovoc against the evaluation framework

Aims and scope. Eurovoc provides a means of indexing documents in the documentation systems of the European institutions, and is a useful search tool for users in general.

Languages and domains. The current 4.2 version of Eurovoc was completed in June 2005 and it is accessible for browsing and searching in 17 EU official languages: Spanish, Czech, Danish, German, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Slovene, Finnish, Swedish (Estonian is currently being revised by the Institute of Estonian Language, and because of lack of translators the Maltese translation is not yet available). Romanian, Bulgarian and Croatian are currently being translated, and the Serbian version is not yet available on the web site of Eurovoc.

Eurovoc is a multilingual polythematic thesaurus covering the fields in which the European Communities are active, for example, politics, law, international relations, employment, agriculture, forestry, fisheries, and energy, among others. The main fields covered by the thesaurus are law and legislation of the European Union (EU).

Steps, sources and techniques used for localizing. Eurovoc working language is French.

- Terms and expressions have been translated by the **Directorate General for Translation**²⁹ (**DGT**) of the European Commission. The main language and translations tools at the disposal of all translation teams that make up the DGT are summarized in Table 3. The general translation workflow is to see in Figure 16.

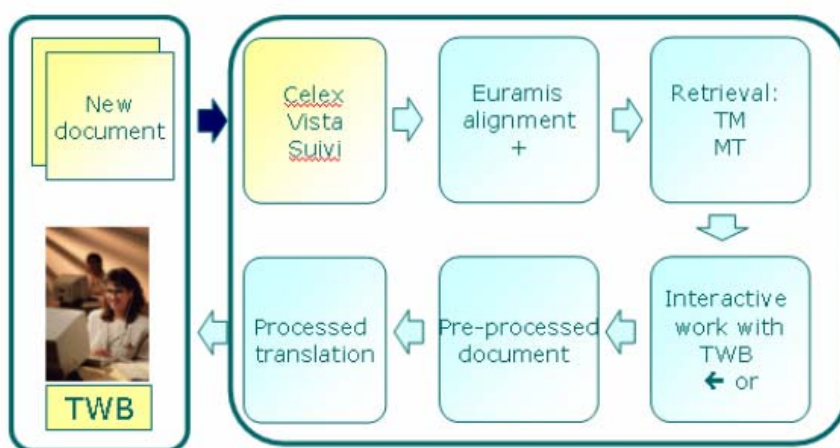


Figure 16: DGT translation workflow (DGT 2005)

Table 3: Main Language and Translation Tools of the DGT

Resource type	Tool name	Function
Language resource	<i>Vista</i>	DGT's electronic archiving system. Contains all original and translated documents from every Directorate General since 1994.
Language resource	<i>Eur-Lex</i>	Repository of the Official Journals of the European Union.
Terminology tool	<i>Eurodicautom</i>	Central terminology database of the European Commission in 11 languages plus Latin.
Translation supporting tool	<i>Quest</i>	Meta-search interface for translators to query several databases simultaneously.
Translation supporting	<i>Euramis Central Translation</i>	Data base layer accessed to retrieve or store data processed

²⁹ http://ec.europa.eu/dgs/translation/index_en.htm

tool	<i>Memory</i>	locally with <i>Trados Translator's Workbench</i> .
Translation supporting tool	<i>Trados Translator's Workbench</i>	Local translation memory connected to the central translation memory <i>Euramis</i> .
Translation tool	<i>EC Systran</i>	A machine translation tool, which actually offers translation for 18 language pairs.
Translation supporting tool	<i>Dragon NaturallySpeaking</i>	Voice recognition tool available for German, Spanish, English, French, Italian and Dutch for supporting the translation task.
Translation supporting tool	<i>Dictatrans (Philips)</i>	Voice recognition tool available for Finnish, Portuguese and Swedish.
Translation supporting tool	<i>TMan</i>	Integrated in the <i>Euramis (European Advanced Multilingual Information System)</i> Web interface, is an automated search-and-replace tool.
Translation supporting tool	<i>Euramis Alignment and Euramis Alignment Editor</i>	Integrated in the Euramis System alignment and alignment editor tools.
Translation supporting tool	<i>Euramis Document Search</i>	Integrated in the Euramis System searcher.
Translation supporting tool	<i>Euramis on-line Concordance</i>	Integrated in the Euramis System concordancer.

- In the framework of the accession of the 10 new countries in 2004, the translation of Eurovoc (version 4.1) has been processed by the national parliament libraries of these countries. Some of these languages have been translated from English, and some discrepancies have been noticed between the basic language (French) and some new translations.

In the following web address of the DGT http://ec.europa.eu/translation/index_en.htm there are links to **publicly available on-line lexical resources**, which are used by the different translation teams of the European Commission and which have been used for the creation of the thesaurus Eurovoc. Some of them are common to all teams; others are specific of each language combination:

- Multilingual encyclopaedias
- Multilingual dictionaries
- Bilingual dictionaries (specific of each language combination)
- Termbanks
- Glossaries

Apart from those lexical resources, other multilingual resources like repositories of EU law in all official languages are at disposal of translators. From all of them, **CELEX** was the most relevant and mostly used for the translation of Eurovoc. CELEX was a repository of EU law in all official languages until January 2005. Users of CELEX have now the option to consult the new **EUR-Lex** website, which incorporates the CELEX database. EUR-Lex provides easy access in 20 languages to the largest documentary database on EU law. It is also possible to view two versions of the same document (mostly original and translation). The system made it also possible to consult the **Official Journal of the European Union**, that includes treaties, legislation, case-law and legislative proposals.

Multilingual information display. In order to look for a term in the Eurovoc thesaurus, we have two options: we can **search** for a term, or **navigate** through the main domains in which the thesaurus is divided (Politics, Law, Finance or Environment, for example). The **language** can be selected in the language menu at the top of the page. At a generic level, Eurovoc has two hierarchical classifications:

- fields, identified by two-digit numbers and titles in words, e.g.:
10 EUROPEAN COMMUNITIES
- microthesauri, identified by four-digit numbers, e.g.:
1011 COMMUNITY LAW

Search: by clicking on the search option, we have the possibility to introduce a term or expression -which can be a descriptor or a non-descriptor- and we will obtain a list of descriptors and non-descriptors that contain the expression entered, as show in

Figure 17. The second step would be to select the searched descriptor in order to obtain its semantic relationships (

- Figure 18), which are:
 - Microthesaurus relationship (abbreviated as **MT**): reference to the field or domain to which the expression belongs.
 - Scope notes relationship (**SN**): definition or usage of the descriptor.
 - Equivalence relationship: relationships between descriptors and non-descriptors shown by the abbreviations **UF** (used for), between the descriptor and the non-descriptor(s) it represents, and **USE**, between a non-descriptor and the descriptor which takes its place. Such relationships are of several types as near-synonymy, antonymy or inclusion.
 - Hierarchical relationship: relationships between a specific descriptor and a more generic one, indicated by the abbreviation **BT** (Broader Term), together with a number showing the hierarchical steps between them; and relationships between a generic descriptor and a more specific descriptor shown by **NT** (Narrower Term), and with the number of steps as well.
 - Associative relationship (**RT**): relationships between two associated descriptors of various kinds, for example, cause and effect, agency or instrument, or location.



Term	Usage
amateur fishing	USE sport fishing
catch of fish	
closed period for fishing	USE fishing season
closed season for fishing	USE fishing season
common fisheries policy	
Community fisheries	
Community fishing	USE Community fisheries
competitive fishing	USE sport fishing
conservation of fish stocks	
crawfish	USE crustacean
crayfish	USE crustacean
deep-sea fishing	
discarded fish	
European Fisheries Guidance Fund	USE FIFG
Financial Instrument for Fisheries Guidance	USE FIFG
fish	
fish croquette	USE fish product
fish disease	

Figure 17: Results for the searched term “fish” in the 1st step of the search



Term	Usage
fish	
MT 5641 fisheries	
UF piscicultural species	
UF species of fish	
BT1 fishery resources	
NT1 fish disease	
NT1 freshwater fish	
NT1 sea fish	
RT fish farming (5641)	
RT fish oil (6016)	
RT fish product (6026)	

Figure 18: Results for the searched term “fish” in the 2nd step of the search

Navigation: the option navigation allows us to obtain all related terms of a specific subject field important for the activities of the European Institutions, e.g. politics, international relations or European Communities, as can be seen in

- . By clicking in on of the descriptors we obtain all related terms together with the relationships between them, as already explained in the **Search** section.

The screenshot shows the Eurovoc Thesaurus interface. At the top, there is a logo for 'eurovoc' and 'THE SAURUS' with a language dropdown menu set to 'English (en)'. The main content is divided into two columns. The left column lists subject fields: '04 POLITICS', '08 INTERNATIONAL RELATIONS', and '10 EUROPEAN COMMUNITIES', each with a list of related terms. The right column shows the microthesauri for the selected term '0406 political framework', including 'political ideology' and 'ecology movement' with their respective relationships (RT) and counts.

04 POLITICS

- 0406 political framework
- 0411 political party
- 0416 electoral procedure and voting
- 0421 parliament
- 0426 parliamentary proceedings
- 0431 politics and public safety
- 0436 executive power and public service

08 INTERNATIONAL RELATIONS

- 0806 international affairs
- 0811 cooperation policy
- 0816 international balance
- 0821 defence

10 EUROPEAN COMMUNITIES

- 1006 Community institutions and European civil service
- 1011 Community law
- 1016 European construction
- 1021 Community finance

0406 political framework

political ideology

- RT political affiliation (0411)
- RT political discrimination (1236)
- RT political party (0411)
- RT political science (3611)
- RT politics (0431)

NT1 anarchism

NT1 Communism

- RT Communist Party (0411)
- RT post-communism (1621)

NT1 conservatism

- RT Conservative Party (0411)

NT1 ecologism

- RT ecology movement (0431)
- RT Ecology Party (0411)

Figure 19: Subject fields and microthesauri of the EC in Eurovoc

Systems of representation of multilingual information. No information has been found referring to this, despite contact established to developers

Evaluation methods. Translations of the Eurovoc terms are manually revised by correctors from the Official Journal Unit of the Publications Office.

URL: <http://europa.eu/eurovoc/>

Accessed September 2006

Contact for information developers: opoce-eurovoc@cec.eu.int

Relevant bibliographic references:

Directorate-General for Translation of the European Commission, July 2005, *Translation Tools and Workflow*, in http://ec.europa.eu/dgs/translation/index_en.htm

9. Lexicon

9.1 EuroWordNet (EWN)

9.1.1 Short description of EWN

EuroWordNet was a 3-year European project (1996 - 1999) that developed a general-purpose multilingual lexicon. This LR was based on and had the same structure of the Princeton WordNet³⁰ (Miller et al. 1990), developed as a monolingual lexical database for American English. Resources and development project were supported by the Human Language Technology sector of the Telematics Applications Programme (Project Reference number: LE-2 4003 & LE-4 8328). In the design of EWN, universities of Holland, Spain, Italy, England, France, Germany, the Czech Republic and Estonia worked together to develop each specific wordnet. For more details on each partner contributors see Vossen 2002.

The work initiated in the EWN project is now being continued by the Global Wordnet Association (GWA)³¹.

9.1.2 Comparison of EWN against the evaluation framework

Aims and scope. The aim of this project was to develop a multilingual lexicon with wordnets for several European languages (see English, Dutch, Spanish and Italian wordnets in Figure 20), which could be used “to improve recall of queries via semantically linked variants in any of these languages”. The general approach for EWN was to build the multilingual database taking advantage of existing resources in each language. Participants from each country were responsible for a language specific wordnet using their already available tools and resources built up in previous national and international projects.

³⁰ <http://wordnet.princeton.edu/>

³¹ <http://www.globalwordnet.org/>

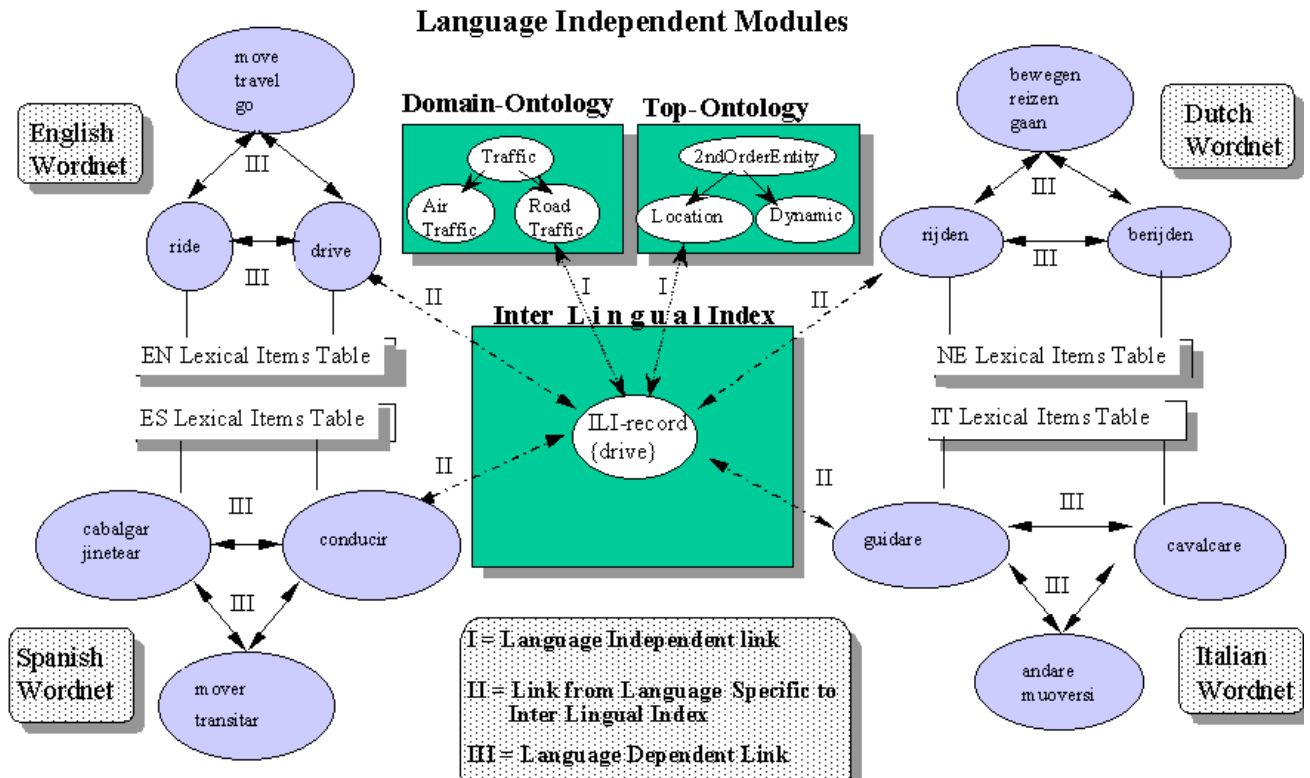


Figure 20: The global architecture of the EWN database (Vossen 2004)

As in WordNet, information about nouns, verbs, adjectives and adverbs was organized in **synsets**. A synset is “a set of words with the same part-of-speech that can be interchanged in a certain context” (Vossen 2004). Synsets are related to each other by semantic relations, such as hyponymy or meronymy, for example.

The **wordnets** in EuroWordNet are considered “autonomous language specific ontologies”. Then, multilingual wordnets are interconnected through an **Inter-Lingual-Index** (ILI), a list of unstructured meanings mainly from Princeton WordNet, specifically WordNet1.5, that provide the mappings across the wordnets, as illustrated in Figure 20. These ILI-records are related to one of more of the 63 top concepts from a **Top Ontology**³² and to domains from the **Domain-Ontology** (also Figure 20). A selection of the ILI-records, the so called **Base Concepts** or **Common Base Concepts**, builds the core of each independent wordnet. BCs have a high position in a hierarchy and able to have links to many other concepts (hyponyms).

Some language-independent structuring is provided to the ILI by the **Top-Ontology** (concepts reflecting semantic distinctions, as Object and Substance, Location, Dynamic, etc.) and the **Domain-Ontology** (topics that group meanings together, as Traffic, Sports, Hospital...).

Languages and domains. As already mentioned, EWN is a general-purpose multilingual lexical database. In the first two phases of the project, wordnets for eight European languages were created: English, Dutch, Italian, Spanish, French, German, Czech and Estonian. The Global Wordnet Association was created in 2000 with the purpose of establishing a word wide association for “maintaining, standardizing and interlinking wordnets for all languages in the world, likewise preparing the ground for the development of a word wide multilingual database with wordnets”.

³² The Top Ontology, created for this purpose, was based on semantic classifications common in linguistic paradigms: Aktionsart models [Vendler 1967, Verkuyl 1972, Dowty 1979, Verkuyl 1989, Pustejovsky 1991, Levin 1993], entity-orders [Lyons 1977], Aristotle’s Qualia-structure [Pustejovsky 1995]; on ontological classifications from previous EC-projects: Aquilex (BRA 3030, 7315), Sift (LE-62030); and was compared with language-neutral ontologies such as CYC, Upper-Model, and Mikrokosmos. For a more detailed information see (Vossen, 2002: 58-71).

Thanks to this association, wordnets have been developed or are being currently developed for other European and non-European languages as: Arabic, Basque, Catalan, Chinese, Danish, Hebrew, Hindi, Korean, Russian, Slovenian, Swedish and Tamil. Moreover, wordnets for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish have been produced within the Balkanet³³ project - or are being maintained by Balkanet as in the case of Czech- a related project to the GWA.

Steps, sources and techniques used for localizing. Bearing in mind the internal structure of EWN, described in the Aims and scope section, we are now to describe how wordnets in each language were developed and expanded from the Base Concepts, which were common to all of them.

BCs were developed to guarantee a minimal level of compatibility between the independent wordnets in each language. In order to expand those core wordnets in each of the EWN languages, two approaches were followed for encoding synsets and semantic relations:

- **Merge model:** synsets and relations are defined separately in a determinate language after which equivalence relations to WordNet1.5 (to the ILI) are generated.
- **Expand model:** WordNet1.5 synsets and relations are translated into equivalent synsets in the other language and are then adapted to EWN, if necessary.

Most of the languages (Dutch, Italian, German, Czech, Bulgarian, Greek, Romanian, etc.) have followed the Merge model in an attempt to maintain the language specific properties. Languages following the Expand model, like Spanish, Basque or French, will result in wordnets “very close to WordNet 1.5, but which can also be biased by it” (Vossen 2002). The top-down extension of the core wordnets, after the selection –or translation- of the BCs, has been done mostly **manually** or using **semi-automatic techniques**, and relying on the information of the adopted resources.

The main resources are:

- monolingual dictionaries
- taxonomies or databases; and
- bilingual dictionaries (English/target language).

In order to obtain a general view of the **steps and techniques used for the localizing** of each wordnet, we have selected two of the independent wordnets –the Dutch Wordnet (Table 4) and the Spanish Wordnet (Table 5)- each one following a different approach.

Table 4: Steps used for localizing the Dutch WordNet

Steps	Development of the Dutch WordNet following the Merge model
1.	Conversion of the Vlis database ³⁴ to the EWN structure and addition of the Dutch lexicon with Celex ³⁵ corpus frequency information.

³³ <http://www.ceid.upatras.gr/Balkanet/>

³⁴ Lexical database provided by Van Dale publisher (Vossen et al. 1999).

Van Dale web at: <http://www.vandale.nl/opzoeken/woordenboek/>

2.	Automatic generation of equivalence relation via the bilingual dictionaries ³⁶ (partly manually and partly with automatic techniques, mapping the Vlis database with bilingual dictionaries, and then mapping the resulting translations to WN 1.5).
3.	Development of the Dutch core wordnet around the Dutch equivalences of the common BCs and other Dutch concepts that are important (taking into consideration the following criteria: number of relations, position in the hierarchy, Vlis top senses and frequency).
4.	Extension of the core wordnet to complete Dutch wordnet.

Table 5: Steps used for localizing the Spanish WordNet

Steps	Development of the Spanish WordNet following the Expand model
1.	Manually mapping of Spanish words to the two highest levels of WN1.5 (BCs) using monolingual and bilingual dictionaries ³⁷ for nouns, and bilingual databases for verbs ³⁸ .
2.	Comparison of that initial set of concepts with BCs sets from other sites of EWN to assure the merging.
3.	Enrichment of the core wordnets with lexical-semantic relations extracted from monolingual dictionaries inexistent in the English language.
4.	Extension of the core wordnet with monolingual dictionaries and semantic taxonomies ³⁹ .

The following figure, Figure 21, shows a global overview of steps in building EWN. Within the production phase (steps 1a and 1b in Figure 21) both **Merge** and **Expand Models** are included in a generalized way.

³⁵ Celex Dutch lemma lexicon. Go to http://www ldc.upenn.edu/Catalog/readme_files/celex.readme.html for a detailed information.

³⁶ Van Dale Dutch-English dictionary (Martin and Tops 1986); Van Dale English-Dutch dictionary (Martin and Tops 1989)

³⁷ DGILE: Diccionario General Ilustrado de la Lengua Española (M.Alvar (ed), Bibliograf. S.A., Barcelona 1987)

English-Spanish and Spanish-English Bilinguals VOX-HARRAP'S Special, and VOX Advanced. (Bibliograf. S.A., Barcelona 1992)

³⁸ PIRAPIDES database (Castellón et al. 1997)

³⁹ Taxonomies developed within the Acquilex project

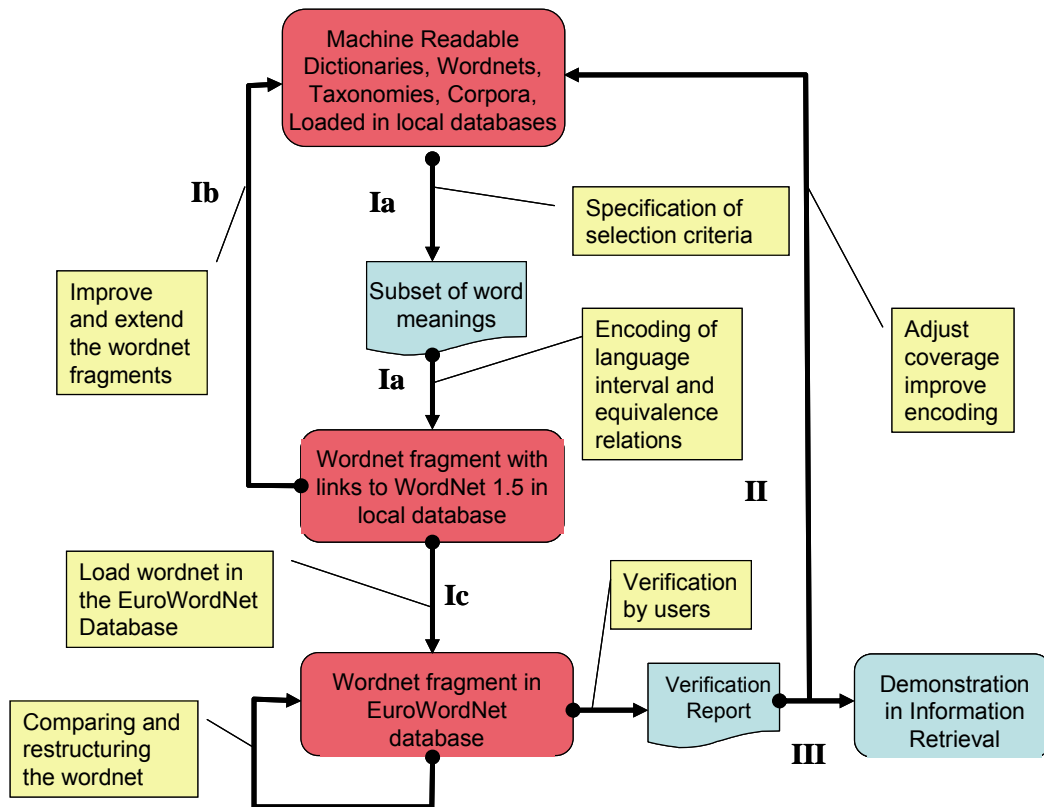


Figure 21: Building steps in EWN (Vossen 2002)

Multilingual information display. The multilingual information is displayed as shown in the example below (Figure 22) from the Interface Meaning 2.0 for the Spanish EWN database. The searched word is displayed accompanied by the corresponding translation in the selected languages. The gloss in English, as well as the score or all possible relations of the word in the database, are optional. On the left side, the base concept and the relations in the Top Ontology are also shown.

fish

Word: Nouns English_1.6

Synonyms: near_synonym English_1.6

Gloss English_1.6 English_1.7 English_1.5

Score Spanish_1.6 English_1.7.1

Rels Catalan_1.6 English_2.0

Full Basque_1.6 Catalan_1.5

Italian_1.6 Spanish_1.5

02005782n

[zoology](#)

base concept 02005782n 609 [fish_1](#) any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills

[animal](#) 02005782n 606 [pez_1](#)

[Fish=](#) 02005782n 606 [peix_1](#)

[Animal=](#) 02005782n 610 [arrain_1](#)

[Living+](#)

[Object=](#)

05810856n 05810856n 29 [fish_2](#)

[gastronomy](#) 05810856n 97 [pescado_1](#) the flesh of fish used as food

[food](#)

[Meat+](#)

[Comestible+](#) 05810856n 97 [peix_2](#)

[Natural+](#) 05810856n 30 [arrain_3](#)

[Substance+](#) 05810856n 31 [pesce_1](#)

07156174n 07156174n 0 [chump_1](#) [fish_3](#) [fool_2](#) [gull_1](#) [mark_8](#) [patsy_1](#) [fall_guy_1](#) [sucker_1](#) [schlemiel_1](#) [shlemiel_1](#) [soft_touch_1](#) [mug_2](#) a person who is gullible and easy to take advantage of

[person](#)

[person](#) 07156174n 0 [berzotas_1](#) [tonto_2](#) [primo_1](#) [panoli_1](#) [mameluco_1](#) [imbecil_1](#) [estúpido_1](#) [cipote_1](#) [bobo_2](#)

[Human+](#) 07156174n 0 [babau_2](#) [tòtil_1](#) [enze_1](#) [encantat_1](#) [gamarús_1](#) [tanoca_1](#) [pallús_1](#) [beneit_2](#)

[Function+](#) 07156174n 0 [gizajo_2](#) [inzente_4](#) [inozo_14](#) [tonto_4](#) [babo_4](#) [txepel_3](#) [ergel_12](#) [kirten_15](#)

[Human+](#)

[Living+](#) 07156174n 0 [allocco_3](#) [babbeo_3](#) [balordo_2](#) [bischero_5](#) [castrone_4](#) [credulone_1](#) [deficiente_2](#) [fanciullone_1](#) [gonzo_1](#) [grullo_1](#) [ingenuo_1](#) [mamaluco_1](#) [mamaluco_2](#) [merlo_3](#) [pagolino_1](#) [paolino_1](#) [pippione_1](#) [pisellone_1](#) [pollastro_2](#) [pollo_2](#) [rimbecillito_1](#) [stolido_1](#) [stolto_2](#) [stupido_1](#) [tonto_3](#) [uccellaccio_1](#)

[Object+](#)

Figure 22: Interface Meaning 2.0

Systems of representation of multilingual information . As can be seen in Figure 23, the 63 Top-Ontology concepts, the 1310 Common Base Concepts (CBC) and the remaining WordNet1.5 Synsets are independent of the specific language wordnets, and form the so-called ILI. To those CBC, complementary Local BCs are added in order to build the core of the language wordnets. CBC and Local BCs are linked to their own specific hyperonyms and hyponyms, and each of these items is again linked to the ILI.

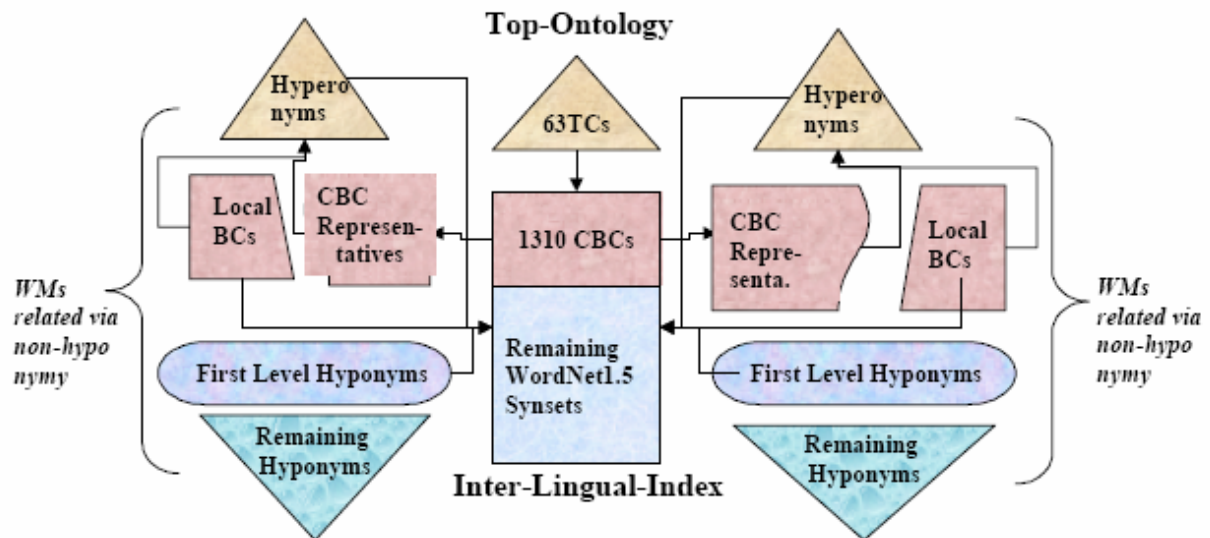


Figure 23: General outline of two wordnets linked to the ILI (Vossen 2002)

Evaluation methods. Evaluation of internal relations in synsets and equivalence relations to the ILI are carried out manually in the III phase of the building process (cf. Figure 21). Verification is done by users, who submit a verification report. Afterwards, a demonstration is done in Information Retrieval (end of phase III).

URL: EWN <http://www.ilic.uva.nl/EuroWordNet/>

GWA <http://www.globalwordnet.org/>

BalkaNet <http://www.ceid.upatras.gr/Balkanet/>

Contact for information developers:

Piek.Vossen@irion.nl or www.vossen.info

Fellbaum@clarity.Princeton.edu

Relevant bibliographic references:

Miller G., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. (1990) (Revised in 1993). "Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography*, 3(4), 235–244.

Martin, W. and Tops, G. A. J. (eds.). (1986). *Van Dale Groot Woordenboek Nederlands-Engels, Van Dale Lexicografie*, Utrecht/Antwerpen.

Rodríguez H, Climent S, Vossen P, Bloksma L, Peters W, Alonge A, et al. (1998) *The top-down strategy for building Eurowordnet: Vocabulary coverage, base concepts and top ontology*. *Computers and the Humanities*; p.117-52.

Vossen, P. (2004) "EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index". *Semi-special issue on multilingual databases (IJL 17/2, June 2004)*.

Vossen, P. (2002) "EuroWordNet General Document". (Version 3, Final, July 1, 2002).

Vossen, P. L, Bloksma and P. Boersma. (1999) *The Dutch Wordnet*, University of Amsterdam.

Vossen, P. (1998) "Introduction to EuroWordNet". In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), *Special Issue on EuroWordNet*. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.

Vossen, P., L. Bloksma, P. Boersma, F. Verdejo, J. Gonzalo, H. Rodríguez, G. Rigau, N. Calzolari, C. Peters, E. Picchi, S. Montemagni, W. Peters. (1998) "EuroWordNet Tools and Resources Report". *EuroWordNet (LE-4003) Deliverable D021D025*, University of Amsterdam.

Vossen, P. (1998) "EuroWordNet: Building a Multilingual Database with Wordnets for European Languages". In: K. Choukri, D. Fry, M. Nilsson (eds), *The ELRA Newsletter*, Vol3, n1, 1998. ISSN: 1026-8200.

Acquilex: <http://www.cl.cam.ac.uk/Research/NL/acquilex/>

ConceptNet: <http://elies.rediris.es/elies2/cap333.htm>

PAROLE :http://www.ub.es/gilcub/SIMPLE/reports/parole/parole_morph/paromor.html

10. Ontology localization approaches

10.1 Termontography approach

10.1.1 Short description of the Termontography approach

Termontography is a method developed to give support to multilingual ontology engineering. This method resulted of the collaboration between terminologists and ontology engineers from the Centrum voor Vaktaal & Communicatie (CVC) at the Erasmushogeschool Brussel, within the framework of the European project FF POIROT (IST 2001-38248). This multidisciplinary approach combines the theories and methods for multilingual terminological analysis of the sociocognitive approach (Temmerman 2000) with methods and guidelines for ontological analysis (Gómez-Pérez *et al.* 1996; Fernández *et al.* 1997; Sure and Studer 2003, cited in Kerremans & Temmerman 2003).

10.1.2 Comparison of the localization approach against the evaluation framework

Aims and scopes. Termontography has been developed for knowledge management and representation of a specific domain, combining field specialist's knowledge and natural language data.

Languages involved. The languages involved in the localization process in the FF POIROT project were English, Dutch, French and Italian.

Steps, sources and techniques used for localizing. Figure 24 shows the six methodological steps or phases that characterise Termontography: (1) analysis, (2) information gathering, (3) search, (4) refinement, (5) verification and (6) validation.

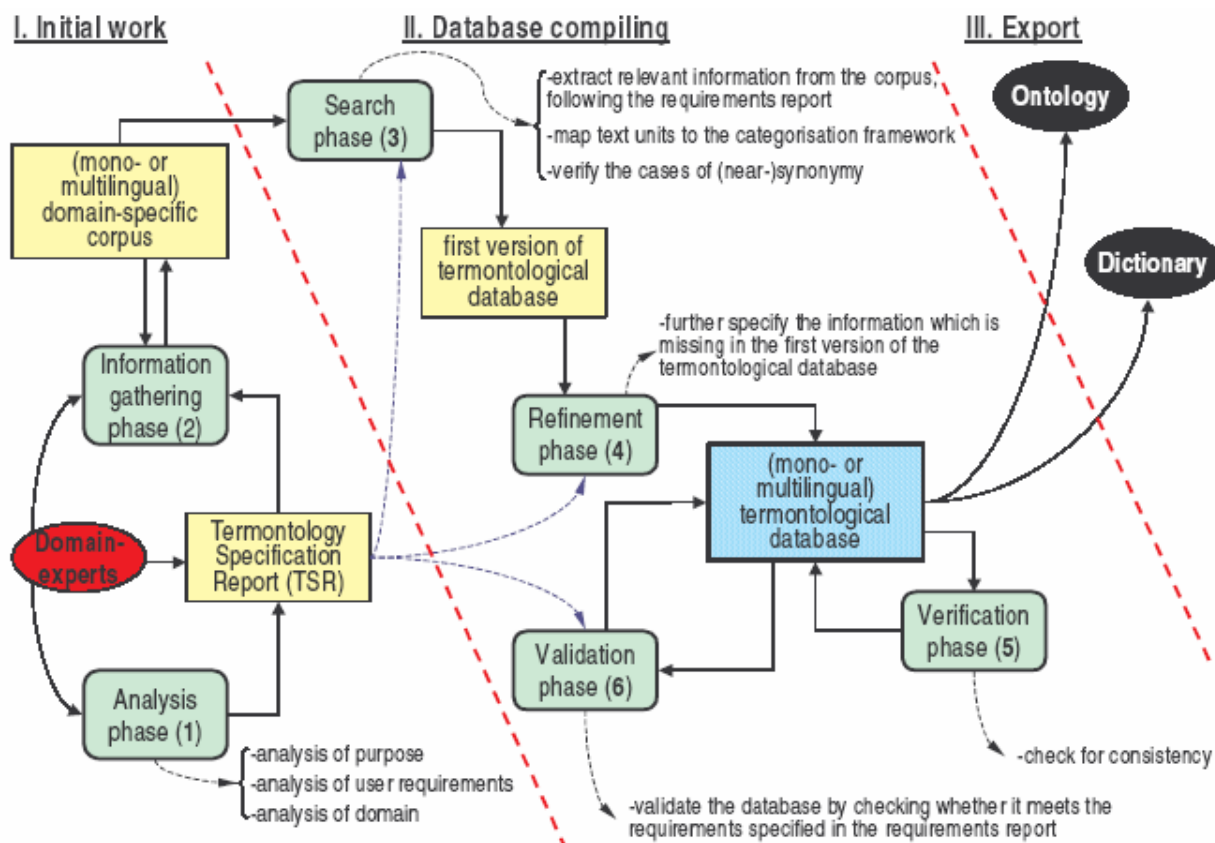


Figure 24: Termonography workflow (Kerremans et al. 2004b)

In the initial work phases (Phases 1 and 2), termonographers work together with experts to determine the scope of the domain, the purpose of the project and the user requirements, which are summarised in the Termontology Specification Report (TSR). In a multilingual project, a multilingual corpus will be compiled for extracting terms and categories for each language separately. In this 2nd phase of Information gathering, as well as in the Search (Phase 3) and Refinement (Phase 4) phases, “**software tools have been used for supporting the process of term extraction and translation**” (Kerremans, Temmerman & Tummers 2004b). These tools, their main function and, developed software are listed in Table 6⁴⁰.

Table 6: Tools for supporting the process of term extraction and translation

Tool	Function	Available software
Web crawler	For automatically retrieving on-line domain-specific-texts	Developed by <i>Knowledge Stones</i> , it retrieves on-line documents based on the clustering of given keywords
Keyword extractor	For giving the user an idea about the content of each document	
Text converter	For saving any electronic format to	

⁴⁰ Note that most of the tools listed above are just prototypes, others are fully operational software systems. However, what is still missing is a common interface that integrates these tools as separate software modules in one workbench.

	plain text	
Automatic aligner	For aligning parallel texts so that only one version needs to be processed during the Termontography search phase	
Similarity measuring tool	For removing one version of two identical documents from the corpus in order to reduce noise for, for instance, the automatic term extractors	
Automatic term identifier	For highlighting, in a new text, the lexicalised units which have already been extracted in previous texts	
Smart concordancer	It indicates important co-text for each term	Provided by <i>Language and Computing nv</i> (Ceusters <i>et al.</i> 2004).
Term extractor	It is able to propose in a new text a list of term candidates, based on the mapping results in the previous texts	Provided by <i>Language and Computing nv</i> (Ceusters <i>et al.</i> 2004). Provided by <i>Knowledge Stones</i> and mainly used for the extraction of Italian terms from Italian domain specific texts.
Translation extractor	It is able to find the translation equivalent of a given term in a bilingual, parallel corpus	TREQ-AL Software, developed by The <i>Research Institute for Artificial Intelligence</i> (described below).

TREQ-AL was used in this project for the extraction of translation equivalents in European directives, starting from a given term list in English (Tufis *et al.* 2003). The predecessor of this system, the so called TREQ, has been already used in word clustering and in checking out the validity of the cross-lingual links between the monolingual wordnets of the multilingual *BalkaNet* lexical ontology (see Stamou *et al.* 2002). The TREQ-AL program takes as input the dictionary created by TREQ and the parallel text to be word aligned. The alignment procedure considers the aligned translation units independent of the other translation units in the parallel corpus. It has 4 steps: left-to-right pre-alignment; right-to-left adjustment to the pre-alignment; determining alignment zones and filtering them out; the word-alignment inside the alignment zones.

Systems of representation of multilingual information. No information has been found referring to this, despite contact established to developers

Evaluation methods. After the search and extraction phases, results are shown in a first version of a termontological database. The user should further manually refine the database by adding or removing information.

URL: <http://cvc.ehb.be/Termontography.htm>

Accessed July 2006

Contact information for developers:

Termontography: {koen.kerremans, rita.temmerman, jose.tummers}@ehb.be

TREQ-AL: {tifis, abarbu, radu}@racai.ro

Language and Computing nv: <http://www.landglobal.com>

Knowledge Stones: <http://www.exprivia.it/AISoftw@re/index.asp>

Relevant bibliographic references:

Ceusters, W., Smith, B. and Fielding, J.M. (2004). "LinkSuite™: Formally Robust Ontology-Based Data and Information Integration", in *Proceedings of DILS 2004 (Data Integration in the Life Sciences)*, (*LectureNotes in Computer Science*), Berlin: Springer.

Kerremans, K. and R., Temmerman. (2004a). "Towards Multilingual, Termontological Support in Ontology Engineering", in *Proceedings Workshop on Terminology, Ontology and Knowledge représentation - 22 & 23/01/2004*, Lyon, France.

Kerremans, K., R. Temmerman and J. Tummers. (2004b). "Discussion on the Requirements for a Workbench supporting Termontography", in *Proceedings Euralex 2004*, Lorient, France.

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. *Proceedings of the International Wordnet Conference*, January 21-25, Mysore, India, 12-14.

http://www.ceid.upatras.gr/Balkanet/pubs/GWA_paper_03.pdf

Temmerman, R. and K. Kerremans. (2003). "Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description", in Hajičová, E., Kotěšovcová, A., Mírovský, J. (eds.), *Proceedings of CIL17*, Matfyzpress, MFF UK (CD-ROM). Prague, Czech Republic.

Tufiş, D., Barbu, A.M., Ion, R. (2003). "TREQ-AL: A word alignment system with limited language resources", *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, May-June, Edmonton, Canada, pp. 36-39.

10.2 LabelTranslator approach

10.2.1 Short description of the LabelTranslator approach

LabelTranslation is a strategy and a platform created for supporting the multilingual extension of ontologies existing in just one natural language. This platform was developed within the European Esperanto project (IST-2001-34373), concluded in 2005, by the Language Technology Lab of the German Research Center for Artificial Intelligence (DFKI GmbH)⁴¹, in Saarbrücken, and the Ontological Engineering Group at the Artificial Intelligence Laboratory of the Universidad Politécnica de Madrid⁴² in Madrid, Spain. Part of this work has been partly continued within the eContent LIRICS project⁴³ (No. 22236), carried out at the IULATERM Group of the Universitat Pompeu Fabra in Barcelona, Spain, from 2004 until 2006.

⁴¹ <http://www.dfki.de/web/>

⁴² <http://www.oeg-upm.net>

⁴³ <http://lirics.loria.fr/>

10.2.2 Comparison of the LabelTranslator against the evaluation framework

Aims and scopes. LabelTranslator was developed in order to support “the supervised translation of ontology labels” (Declerck *et al.* 2006) and, at the same time, to allow for the semantic annotation of multilingual web documents using the resulting multilingual labels of ontologies. By “supervised translation”, it is meant that this approach foresees the intervention of the domain expert or translator in case no results outcome, or for validating them. Therefore, LabelTranslator offers a semi-automatic strategy. LabelTranslator can be integrated into any ontology engineering platform to enable its users to translate their ontologies inside the application.

Languages involved. LabelTranslator is available for Spanish, English and German.

Steps, sources and techniques used for localizing. For the development of LabelTranslator already available multilingual semantic resources and basic natural language processing tools were reused for providing a semi-automatic translation of labels in ontologies. In the current version of the LabelTranslator platform three types of multilingual resources are included:

- EuroWordNet (EWN)⁴⁴, a semantic lexical resource.
- Wikipedia⁴⁵, the multilingual free encyclopaedia on the Web, based on knowledge of the word.
- BabelFish⁴⁶, an on-line translation service used as “fallback position” (Declerck *et al.* 2006).

The steps for the localizing approach are summarized in Table 7:

Table 7: Steps, sources and techniques used for localizing in LabelTranslator

Steps	Sources and techniques
1.	Upload of an ontology in the LabelTranslator platform
2.	Selection of the ontology labels to be translated in one of the target languages (en, es, de)
3.	The system accesses the EWN database for finding the selected term (or part of a term), and also checks in the WordNet database, only if the source language is English
4.	Result(s) (synset and gloss) are displayed, if the matching is successful. Users can then validate the suggestions, modify the translation and save it in the database. A disambiguation problem can as well occur (see Disambiguation problem below)
5.	If the matching in EWN is not successful, the system checks in Wikipedia, which also uses a mechanism for relating entries in the various languages available
6.	If steps 3. and 5. do not provide any results, the system turns to BabelFish
7.	If the translation is still not satisfactory, the user can enter a translation, together with part-of-speech information and a definition

If the same translation session is repeated in the future, the system will return the translation already saved in the memory.

Developers of LabelTranslator give priority to the EWN resource because a “high quality in the translation is expected since “EWN has been built following semantic considerations and validated by language and/or domain experts” (Declerck *et al.* 2006).

⁴⁴ See section 7.1 of the present document for a detailed description of the approach

⁴⁵ <http://es.wikipedia.org/wiki/Wikipedia>

⁴⁶ <http://babelfish.altavista.com/>

Disambiguation problem. In the translation step using EWN (**step 3.**), sometimes more than just one result (or synset) is returned, which could be the appropriate equivalent translation for the label in the ontology. Then, glosses offered by EWN can be of great help, since the system can use them for **disambiguating**. Two approaches -or a combination of both- can be used, and these are the following (Note that LabelTranslator developers suggest the implementation of a hybrid approach combining both strategies):

- **Rule-based strategy:** the terms in the gloss of the target language are also present in the ontology; source and target languages share the same or similar glosses.
- **Static strategy:** based on two gloss-based similarity measure algorithms used in the *Perl package WordNet::Similarity*.

In order to solve the disambiguation problem in Wikipedia (**step 5.**), the user can go to the Wikipedia encyclopaedic articles and manually check that the content, context, etc. of a term match with the ontology content.

Multilingual information display. Figure 25 shows the interface of LabelTranslator. On the left side, uploaded ontology and selected languages are displayed. On the right side, translation options and glosses are offered after the system has checked in EuroWordNet and Babelfish.

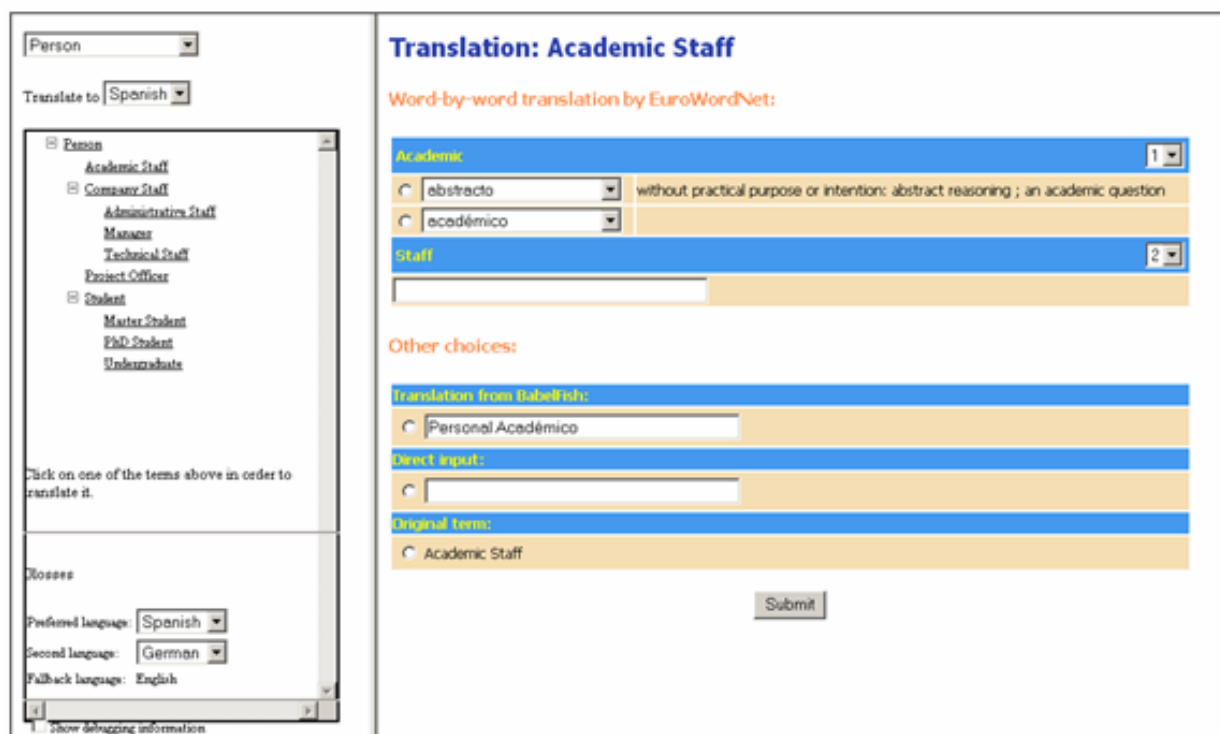


Figure 25: LabelTranslator interface

Systems of representation of multilingual information. The following description is related to the access of LabelTranslator to the multilingual data –in the linguistic resources EWN and Wikipedia, and in the machine translation resource Babelfish–, and to the storage of that information in the LabelTranslator system. Since this is a supporting tool, the final storage of the linguistic information will take place at the knowledge representation level, i.e., in the ontology. The

E/R Schema will then depend on the representation schema of the resource undergoing the localization process.

However, as already said, we will report on the access of LabelTranslator to EWN and the storage of multilingual information during the translation process. For the task of querying the data within EWN, and the retrieval of translations from it, a Java API (see Figure 26) was created (Gantner 2004). The EWN data is stored in distinct MySQL databases. All databases have the same schema and can be accessed by the same SQL statements, which are contained in the monolingual API. The multilingual API consists of several objects of the monolingual API (one for each language i.e., currently English, German and Spanish), and a routine to get translations from and to any of the mentioned languages.

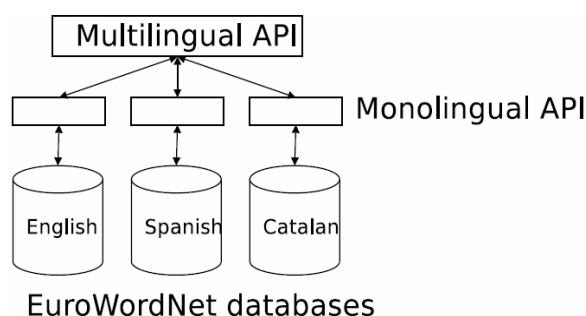


Figure 26: API structure (Gantner 2004)

Evaluation methods. The user can check whether results submitted by EWN are appropriate, and then compare them with results by Babelfish or Wikipedia. In Wikipedia the user has the option of reading the encyclopaedia information related to the term and check if the contextual information corresponds to the selected one.

Contact for information developers.

zq@zenogantner.de

declerck@dfki.de

asun@fi.upm.es

Relevant bibliographic references:

Declerck, T., A. Gómez-Pérez, O. Vela, Z. Gantner, D. Manzano-Macho (2006). "Multilingual Lexical Semantic Resources for Ontology Translation". *Proceedings of LREC 2006*.

Declerck, T. and O. Vela (2005). "LabelTranslator: Multilingualism in Ontologies". *Proceedings of the 4th International Semantic Web Conference. 2005*.

Gantner, Z. (2004). *TermTranslation – A Tool for the Semiautomatic Translation of Ontologies*. Technical report written at the Ontology Engineering Group of the UPM, Spain [unpublished].

Data accessed in September 2006

10.3 OntoLing Tab approach

10.3.1 Short description of the OntoLing approach

OntoLing is a framework for a semi-automatic linguistic enrichment of ontologies. It has been developed at the AI Research Group, Department of Computer Science, Systems and Production of the University of Rome, Tor Vergata. Armando Stellato is the person in charge of its development. OntoLing has been designed as a plug-in for Protégé⁴⁷, a popular ontology editor developed by Stanford Medical Informatics at the Stanford University School of Medicine, allowing the linguistic enrichment of ontologies created within this working environment. The last update of the tool in October 2006 is available under <http://ai-nlp.info.uniroma2.it/software/OntoLing/>, and can be downloaded for free.

10.3.2 Comparison of the OntoLing against the evaluation framework

Aims and scopes. The OntoLing framework was developed for “supporting manual annotation of ontological data with information from different, heterogeneous linguistic resources” (Pazienza & Stellato 2006a). The latest version of OntoLing even helps the user with automatic suggestions through the exploitation of different linguistic resources. By exploiting existing bilingual resources, OntoLing helps in the development of multilingual ontologies, “in which different multilingual expressions coexist and share the same ontological knowledge” (*ibidem*). In this sense, if ontologies are already available in one natural language, this tool helps in the process of ontology localization or, as has been defined by its developers, in the “multilingual enrichment process” (*ibidem*).

Languages and domains involved in the localization process. In the current version of OntoLing, two LRs are available for the linguistic or multilingual enrichment, WordNet⁴⁸, for the linguistic enrichment of ontologies with English labels, and DICT dictionaries⁴⁹, for the linguistic and multilingual enrichment of ontologies (see Figure 27). This last resource accesses a compendium of multiple on-line monolingual and bilingual dictionaries, as for example, all bilingual Freedict Dictionaries: English-German, English-Arabic, English-Croatian, English-Hungarian, etc.

⁴⁷ <http://protege.stanford.edu/>

⁴⁸ <http://wordnet.princeton.edu/perl/webwn>

⁴⁹ <http://www.dict.org/links.html>

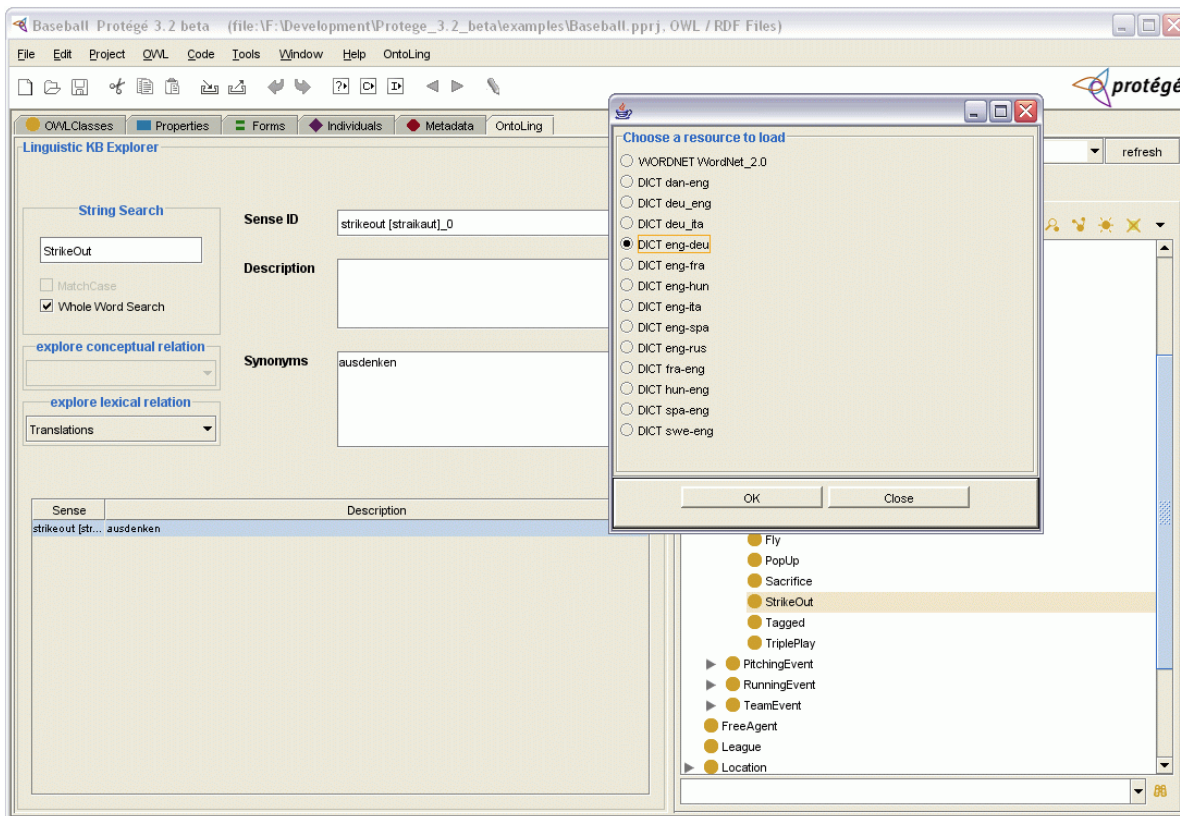


Figure 27: Selection of LRs in OntoLing

Steps, sources and techniques used for localizing. Since OntoLing has been developed as a plug-in for Protégé, the user has to upload an ontology in the Protégé ontology editor in order to use it. Any Protégé plug-in, exploiting linguistic resources, includes a linguistic watermark package, i.e., a package that contains abstract classes and interfaces for accessing linguistic resources. As already mentioned, the current package contains two implemented linguistic interfaces related to freely available resources, namely: WordNet and DICT dictionaries.

Steps and techniques of this localizing tool have been summarized in Table 8.

Table 8: Steps, sources and techniques used for localizing in OntoLing

Steps	Sources and techniques
1.	Open an ontology in the Ontology Panel of the Protégé editor.
2.	Select from the OntoLing menu of available linguistic resources those that will be visualized during the translation task (see Figure 27).
3.	OntoLing accesses the selected linguistic resources by means of a wrapper called Linguistic Interface. With this Linguistic Interface the user visualizes the linguistic information in the Linguistic Browser Panel embedded in the Protégé framework, as Figure 28 shows (left hand side of the OntoLing panel).
4.	The ontology can be enriched with: <ul style="list-style-type: none"> • Additional labels for the selected class, i.e., synonyms • Glosses as descriptions for the selected class • IDs of the selected senses as additional labels for the selected class. This is useful if pointers from ontology concepts to senses from a given LR are needed.
5.	The user checks the suggestions offered by the linguistic enrichment module and selects the appropriate ones (see Figure 28).

6.	Selections are added to the ontology.
----	---------------------------------------

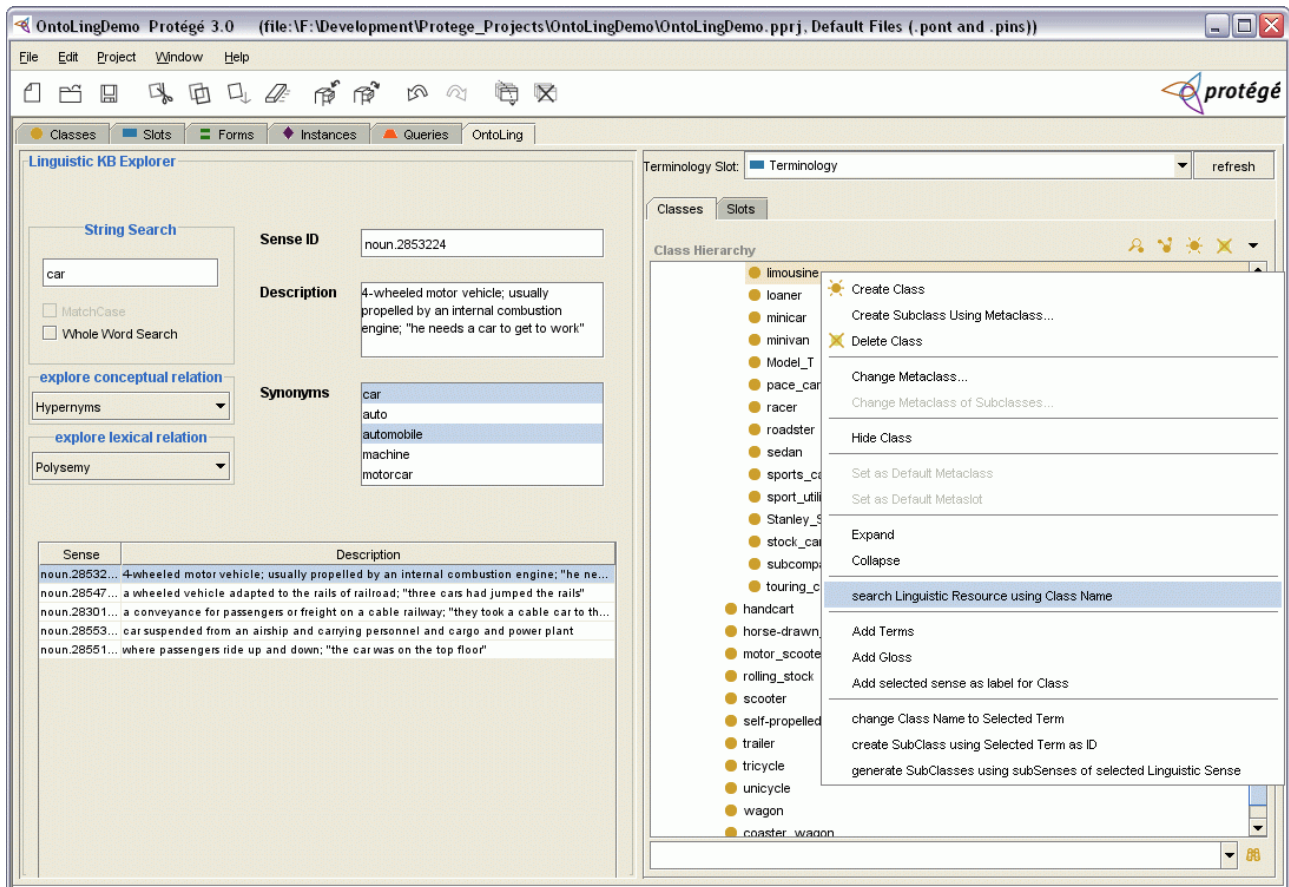


Figure 28: Linguistic Browser Panel in OntoLing

Regarding the **automatic linguistic enrichment of ontologies**, this is currently under development. Moreover, this functionality will be only available if the ontology is in OWL (Web Ontology Language)⁵⁰, and the loaded linguistic resource is a taxonomical lexical resource and/or a linguistic resource with glosses. The enrichment component will exploit the taxonomical structure of the glosses of the linguistic resource to judge which linguistic information can be used to enrich the ontology.

Multilingual information display. As we have seen in Figure 27 and Figure 28, OntoLing allows for the selection of linguistic information to enrich the ontology or to translate ontology labels. Up to now, only the following linguistic information can be added to the ontology:

- Sense ID
- Description or gloss
- Set of synonyms

System of representation of multilingual information. Linguistic and multilingual information obtained from the different linguistic resources used to enrich the ontology will be stored in the ontology itself. If the ontology already exists, its meta-model will have to undergo modifications in

⁵⁰ <http://www.w3.org/TR/owl-features/>

order to hold the linguistic information added to the concepts of the ontology. If a multilingual ontology is being developed from scratch, the ontology meta-model will have to introduce new properties to describe ontology classes: properties containing linguistic information. For those ontologies developed in OWL, these properties are set by default in the `rdfs:label` property and the `owl:comment` property for definitions or glosses (as illustrated in Figure 29).

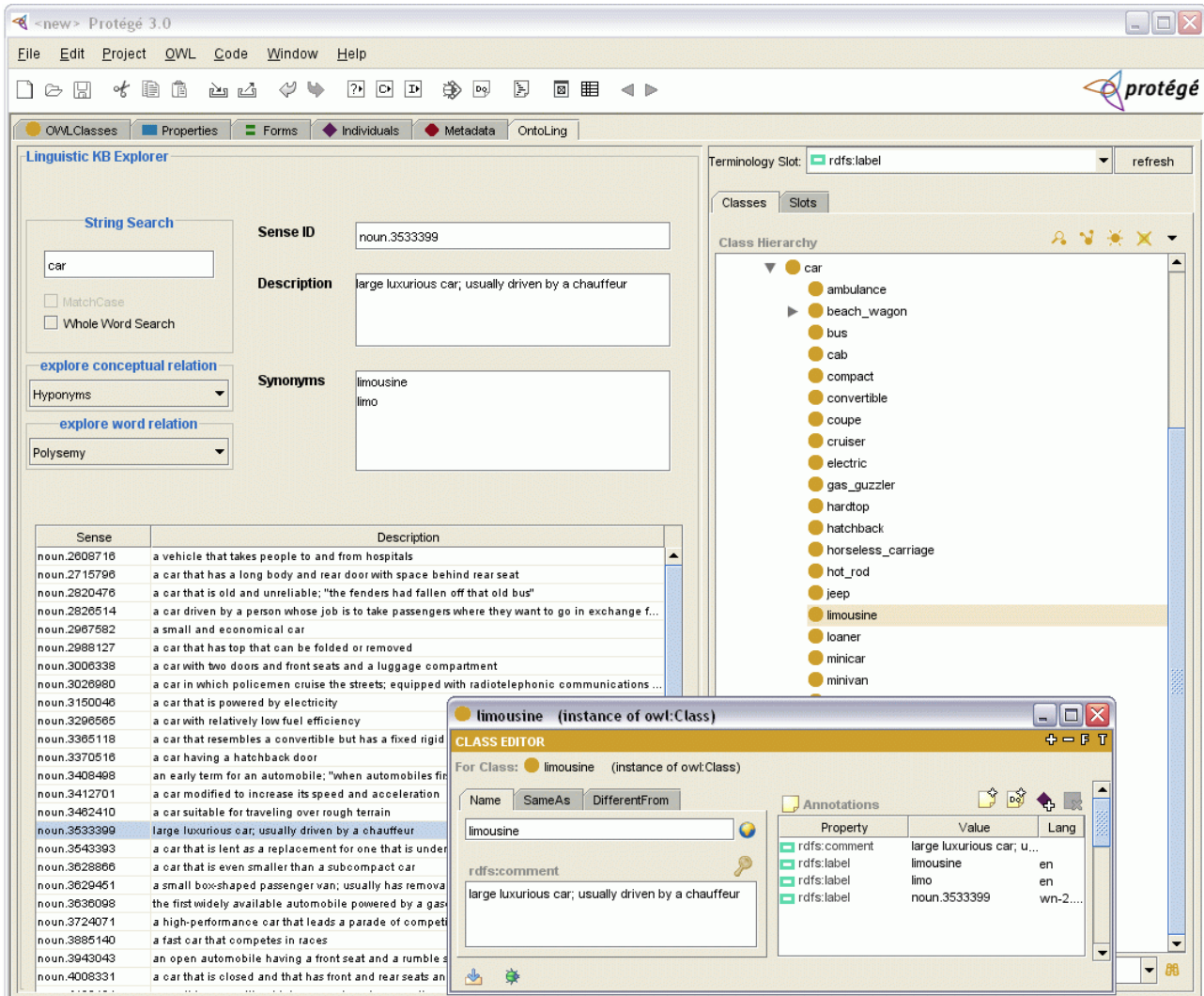


Figure 29: Inclusion of multilingual data in ontologies in OWL

Evaluation methods. The evaluation has to be carried out manually. The translator, terminologist or expert in charge of the linguistic aspect of the ontology will be ultimately responsible for making the right choice when selecting the synonyms, descriptions, and so on, for the ontology concepts.

URL: <http://ai-nlp.info.uniroma2.it/software/OntoLing/>

Accessed February 2007.

10.4 GENOMA-KB approach

10.4.1 Short description of GENOMA-KB

The Human Genome Knowledge Base (GENOMA-KB) is an ongoing research project started in 2001 at the Institute of Applied Linguistics (IULA) of the Universitat Pompeu Fabra in Barcelona, Spain. The IULATerm group research is in charge of this project, whose main objective is the construction of a biomedical knowledge base for the human genome. This research was carried on within the framework of two public funded projects: *TEXTERM: Textos especializados y terminología: selección y recuperación automática de la información* (BFF2000-0841), led by M.T. Cabré; and *RICOTERM: Sistema de recuperación de información con control terminológico y discursivo* (TIC2000-1191), led by M. Lorente. The TEXTERM Project aimed to provide a theoretical basis for computer-aided unit detection and semiautomatic mapping of cognitive nodes and conceptual relations. The main objective of RICOTERM was to build an IR system capable of improving current systems using terminological control. Both projects finished in 2003 and are currently being continued in a second phase in TEXTERM-2⁵¹ and RICOTERM-2.

10.4.2 Comparison of GENOMA-KB against the evaluation framework

Aims and scope of the localization approach. This project aims to become an essential resource for information retrieval with terminological control in the field of the human genome. The resulting set of knowledge can be used for different tasks, such as, document indexation and summarization, machine translation support, etc. Main target users of this LR are, according to the authors, translators, terminologists and lexicographers; information science experts; specialized writers and journalists; researchers and scholars; linguists.

Languages and domains involved in the localization process. The languages involved in this project are English, Spanish and Catalan, and the field of study is the human genome domain.

Steps, sources and techniques used for localizing. In order to understand the localization process of this LR, we need to describe the architecture of the knowledge base. As shown in Figure 30, the knowledge base is divided in four interrelated modules: ontology module, term base module, corpus module and entities module.

- **Ontology module:** the ontology module was developed following the Mikrokosmos design adopted by OntoTerm®⁵², because this terminological management tool allowed the construction of the ontology, integrating at the same time the ontology and the terminological database. This tool provided a core ontology with the 21 basic concepts from Mikrokosmos⁵³ (ALL, OBJECT, EVENT, PROPERTY, etc.). A list of 100 concepts was then added to the initial ones, which were proposed by experts in the human genome domain. The rest of the concepts were recovered from textual specialized information with the aid of lexical resources. In the Ontology Editor a brief description and the conceptual relation was introduced for each new concept. Concepts were fully described with the use of conceptual relations, properties and the inherited information from parent concepts. Possible conceptual relations are (for a more detailed description of the conceptual relations see also Feliu 2004):
 - Similarity
 - Hyponymy
 - Place and time sequenciality
 - Causality

⁵¹ <http://texterm.iula.upf.edu/2/index.html>

⁵² <http://www.ontoterm.com/>, developed by Antonio Moreno at the University of Málaga. For more information see also Moreno et al. 2000

⁵³ <http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html>

- Instrumentality
 - Meronymy
 - Association
- **Term base module:** the information given for each term was the concept expressed by the term. No term entry was possible if the corresponding concept had not been previously introduced in the Ontology module. The information in the Term base was:
 - The term in Catalan, Spanish and English
 - Part of speech
 - Number and gender
 - Usage contexts and its sources
 - The lemmatised form and administrative data
 - **Corpus module:** text corpus of the genomic domain selected and validated by experts, and processed using NLP applications. Texts were in Catalan, Spanish and English.
 - **Entities module:** This module was organized in two parts:
 - Bibliographic module: compiles full references of the information sources used in the Term base and the Corpus base. The languages of the references are also Catalan, Spanish and English.
 - Factographic module: collects updated data about relevant research centres, people, institutions, etc.

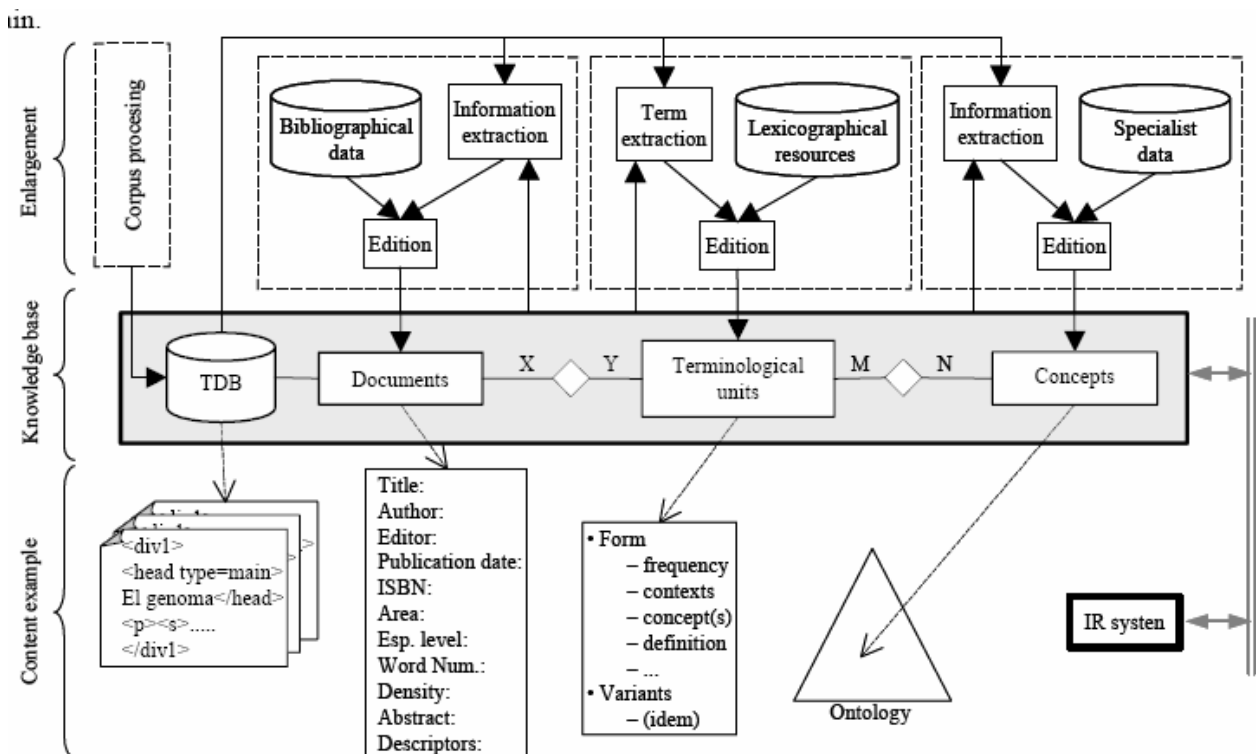






Figure 30: Knowledge Base architecture (Feliu et al. 2002)

The localization process can be summarized in the following steps.

Table 9: Steps, sources and techniques in the localization of GENOMA-KB

	Steps	Sources and techniques
1.	Development of the Ontological module based on ontology concepts and its relations	Mikrokosmos OntoTerm®
2.	Representation of the ontology labels in English	
3.	Compilation of the Corpus module with genomic domain documents selected and validated by experts, in Catalan, Spanish and English.	300 Articles 80 Monographs 30 Specialized journals 30 Ph dissertations
4.	Development of the Term base module in Catalan, Spanish and English, which consists of specialized knowledge units extracted from the specialized corpora (Corpus module) and from on-line dictionaries (or other lexical resources). The extracted terms are then mapped onto the ontology	OntoTerm® <i>Diccionari Enciclopèdic de Medicina d'Enciclopèdia Catalana for Catalan;...</i>
5.	Up to three contexts for each term are included in the Term base module	From the Corpus module , when available; otherwise from Internet.
6.	Non-mandatory definitions from specialized dictionaries are added	
7.	The full bibliographical data is located in the Entities module	

How multilingual information is displayed. The web page of the GENOMA-KB offers the user multiple search possibilities. One can choose to consult amongst the Ontology/Term base module, the Corpus module, the Bibliographic module or the Factographic module. For the purposes of our study, the Ontology/Term base module is going to be the most interesting search. Results for the term “cell” in the Ontology are grouped in 4 sections:

-  Hyperonymy relations
-  Hyponymy relations
-  Co-hyponymy relations
-  Other relations, which can in turn be “is component of”, “is whole component of”, “is place of”, “is whole area of”, “is generally associated with”, “is located in” and “locates”.

Not all relations are displayed at once, but one has to look for each kind of relations at a time. In Figure 31 we see how hyperonymy relations for the term “cell” are presented. Next to the term “cell” there are links to the other types of relations, which can easily be consulted.

The screenshot shows the GENOMA-KB interface. At the top left is the logo 'genoma' and 'Banc de Coneixement sobre el Genoma Humà'. Below it is 'BT banc terminològic'. On the right, there are logos for 'iULA Term' and 'UNIVERSITAT DE VALÈNCIA'. The search bar contains 'Terme de la cerca: cell (Anglès)' and 'Condicció de cerca: Lema exacte'. Below the search bar, it says 'Resultat de la cerca' and 'Termes trobats (1)'. The main content area shows 'cell (CELL)' with a list of hyperonyms: 'All', 'OBJECT', 'PHYSICAL-OBJECT', 'NATURAL-OBJECT', 'ORGANIC-STRUCTURE', and 'CELL'. The 'CELL' term is highlighted with a red box and a red arrow pointing to it from 'ORGANIC-STRUCTURE'. The background of the interface is a light blue pattern of various words.

Figure 31: Hyperonymy relations for the term “cell” in GENOMA-KB

Systems of representation of multilingual information. The GENOMA-KB is built upon four independent modules, as shown in Figure 30. The architecture that supports this platform is the following for each module:

- **Ontology module:** is formed by an Ontological database (Microsoft Access Database), which has been built using the terminology management system OntoTerm®. The ontology contains the concepts and the relations between concepts.
- **Term base module:** based on a Terminological database, Genoterm, (Microsoft Access Database), also built using OntoTerm®. Genoterm is intimately interrelated with the Ontological database and the predefined concepts are associated to the terms.
- **Corpus module:** uses the Corpus Work Bench®⁵⁴, a workbench for full-text retrieval from large corpora. This workbench is used for the extraction of linguistic knowledge, evidence for lexical descriptions, and terms. It also includes an interface for making queries, the bwanaNet⁵⁵, which in turn uses the Corpus WorkBench® tool CQP (Corpus Query Processor) to query the Corpus itself.
- **Entities module:** consists of bibliographical data (Microsoft Access Database) for documents that form part of the corpus, and of factographical data related to people, institutions, etc.

⁵⁴ <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>, developed by IMS of Stuttgart University

⁵⁵ <http://bwananet.iula.upf.edu/indexen.htm>, developed by IULA

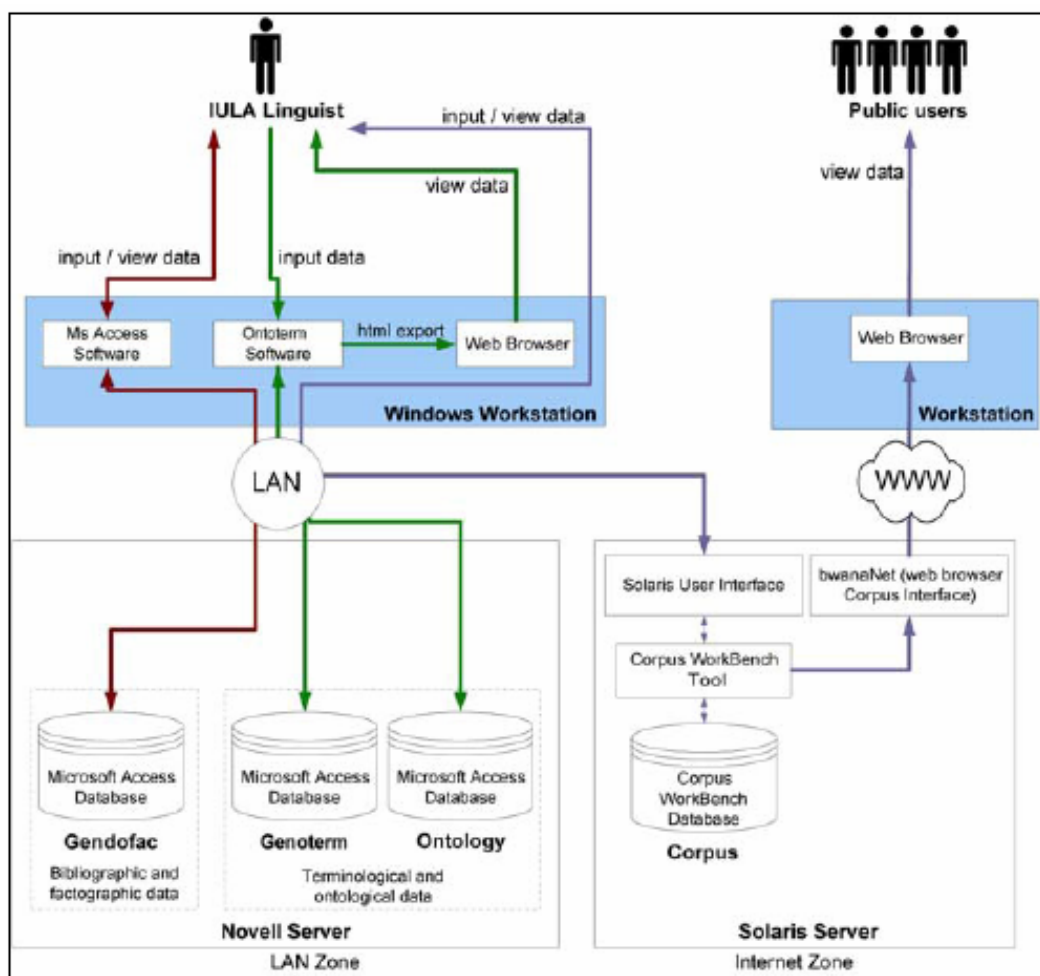


Figure 32: GENOMA-KB architecture support (Hospedales y Rodríguez 2004)

Evaluation methods. Linguists, terminologist and translators of the IULATerm research group are responsible for the evaluation of knowledge base, which will be carried out manually. Public users can also make suggestions.

URL:

<http://genoma.iula.upf.edu:8080/genoma/corpSearch.do;jsessionid=C5F6DA7C2954A5084D48F35666F8B0DE?operation=init>

Contact for information developers:

{teresa.cabre; carme.bach; rosa.estopa; judit.feliu; gemma.martinez; jorge.vivaldi} @upf.edu

{mhospedales; mrodriguez} @spoc.com

Antonio Moreno: amo@uma.es

Relevant bibliographic references

Cabr , M. Teresa; Bach, C.; Estop , R.; Feliu, J.; Mart nez, G.; Vivaldi, J. (2004a). "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities". *LREC 2004 Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Languages Resources Association. pp. 87-90.

Cabré, M. Teresa; Estopà, R.; Feliu, J. (2004b). "A Specialized Knowledge Base: from Distributed Information to the Specialized Dictionary Construction". *11th EURALEX International Conference Proceedings*. Lorient: Euralex. pp. 867-872

Feliu, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Barcelona: Institute for Applied Linguistics. [Ph. Dissertation]

Feliu, J; Vivaldi, J.; Cabré, M.T. (2002) "Towards an Ontology for a Human Genome Knowledge Base". *LREC2002. Third International Conference on Language Resources and Evaluation. Proceedings*. Las Palmas de Gran Canaria, pp. 1885-1890. ISBN: 295-1740-808.

Hospedales, M.; Rodríguez, M. (2004) "The GENOMA-KB platform: Queries over integrated of linguistic resources" *LREC 2004 Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Languages Resources Association.

Moreno Ortiz A. Y Pérez Hernández (2000). "Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases". *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, pp. 1061-1067.

10.5 OncoTerm approach

10.5.1 Short description of the OncoTerm approach

OncoTerm is the name of a research group and a project carried out by interdisciplinary researchers from the Universities of Granada, Málaga and Valladolid, and collaborators from the Hospital Virgen de las Nieves in Granada, in which a multilingual information system on oncology was developed from 1999 until 2002. This project received funding from the Spanish Ministry of Education and Culture. The result of this research is an ontology of 1896 concepts from the oncology field, and a database related to the ontology with more than 4000 terms in Spanish and English.

10.5.2 Comparison of the OncoTerm against the evaluation framework

Aims and scopes. The aim of this project was to create a complete terminological database in oncology, useful for different target users such as health experts, researchers, patients and families, as well as translators and authors of specialised texts. As developers of this ontological resource state, this tool aims at making work more efficient for users, as "it guarantees quality and the access to the requested information in shorter time, thanks to the search system and the concepts organization" (Crónica Universia).

Languages and domains involved in the localization process. Languages involved in this project are Spanish and English, and oncology is the main field of study.

Steps, sources and techniques used for localizing. As in the case of GENOMA-KB (analyzed in this document in **section 10.4**), the OncoTerm terminology database is built upon the OncoTerm™ terminology management system. The OncoTerm™ database consists of two fundamental modules, namely, the **Ontology Editor** and the **TermBase Editor** (cf. Figure 33).

In the first place, a corpus of validated texts by experts and terminologists is compiled and stored in a **Corpus module**. Concepts and terms are then extracted from this corpus, and references to it are made explicit in the terminology database. The **Ontology Editor** organizes the oncology concepts in an ontology that takes as its basis the upper level ontology Mikrokosmos. Mikrokosmos aims at organizing general knowledge in a way which is independent of any language, by classifying the knowledge of the world, i.e., all entities into *objects*, *events*, and

properties. The OncoTerm tool provides 21 preestablished core nodes which have been derived from the implementation of the basic top nodes of Mikrokosmos. The next step consists in linking the specialized knowledge of the oncology field to the upper level ontology. The **TermBase Editor** is in charge of the terminological data. It is in this module where the conceptual model designed in the Ontology Editor acquires its linguistic dimension. It is worth stressing the importance of the OncoTerm™ structure, to show how a new term cannot be introduced if the corresponding concept has not been previously included in the Ontology Editor.

The design steps followed in this resource are very close to the GENOMA-KB development.

Table 10: Steps, sources and techniques in the localization of OncoTerm

	Steps	Sources and techniques
1.	Compilation of a Corpus module with medical domain documents selected and validated by experts, in Spanish and English, as well as interviews with specialists. Texts in paper format are computerized for the purpose of accelerating the process of terms recovery.	<p>Sources: Previous research projects in similar or related subject fields. Documents from the Department of Oncology and Radiology at the Hospital Virgen de las Nieves in Granada, as well as interviews with specialists from the mentioned Hospital.</p> <p>Internet web pages from cancer international organizations, encyclopaedias, medical guides and papers related to cancer⁵⁶.</p>
2.	Extraction of frequency words from the Corpus module taking as starting point concordance lines.	Techniques: <i>Wordsmith Tools</i> ⁵⁷ (a concordancer that helps in text analysis and terms recovery).
3.	Development of the Ontological database consisting of ontology concepts and its relations. Experts and terminologists assign a place for the concept in the ontology model and establish the relations of that concept to the rest of concepts. Starting point for the ontology development is the	Mikrokosmos Ontology ⁵⁸ OntoTerm™

⁵⁶ Cancer international organizations: *CancerNet, CancerBacup, Medscape, MedicineNet, Oncoweb, Virtual Hospital, Alcace, Atheneum y Diario Médico*.

Spanish texts published in *Medicina Clínica, Revista Clínica Española, Neoplasia, Revisiones en Cáncer, Revista Española de Anestesiología y Reanimación, Archivos Bronconeumológicos, Revista Española de Enfermedades Digestivas, Anales Otorrinolaringológicos Ibero-Americanos, Anales Españoles de Pediatría y Actas Urológicas españolas*.

English texts published in *British Medical Journal, Lancet, New England Journal of Medicine, Cancer, CANCERLIT, C-A. A Cancer Journal for Clinicians*.

Medical guides: *Harrison's Principles of Internal Medicine, Cancer: Principles and Practice of Oncology, Medicina Interna de Farreras-Rozmán, Cancer. Principios y Práctica de Oncología y Oncología Médica-Guía de Oncología Médica*.

Encyclopaedias: *The Merck Manual of Diagnosis and Therapy / Manual Merck en español y Mosby's Medical Encyclopedia for Health Consumers*.

⁵⁷ Available under http://www.lexically.net/downloads/version4/wordsmith_versions.htm

⁵⁸ Available under <http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html>

	Mikrokosmos Ontology.	
3.	Representation of the ontology labels in English (that usually do not have a total correspondence with the terms in the English language in the TermBase).	
4.	Development of the TermBase in Spanish and English, which consists of specialized knowledge units extracted from the specialized corpora (Corpus module). The extracted terms are then mapped onto the concepts in the Ontological database .	OntoTerm™
5.	Administrative data related to the concept are included in the TermBase , as for example, the name of the originator of the concept and URL where it has been extracted.	From the Corpus module , and from other terminological resources and LRs in general.
6.	Linguistic data are added to the TermBase : terms in both languages English and Spanish, as well as definition, part of speech, number, term type, and context. Administrative data are also added to the terms, namely, origination date and originator. Images are as well included next to the terminological data, because they offer some helps to the end user.	Use of CLS Framework ⁵⁹ and Reltef™ ⁶⁰ for modelling and storing the information in the term base (explained in the Systems of representation of multilingual information section).

How multilingual information is displayed. The OncoTerm web page is divided in three sections:

- Introduction to the Terminological Database OncoTerm and information about developers
- Description of the project
- Access to the terminological database OncoTerm

The first two sections are written in Spanish, the terminological database is bilingual in English and Spanish.

When clicking on the terminological database link, the user accesses the alphabetical list of concepts of the ontological domain placed on the left hand side, as can be seen in Figure 33. By selecting one of the concepts, terminological information in English and Spanish is displayed on the right hand side. Fix text parts are only in English. A kind of “global information” section appears at the beginning on this terminological part including information about the “subject field”, “origination date”, “originator” and “URL” of the concept. Then, a table gathers information about the ontological relations of that concept within the ontology. Relations included in this table are: hierarchical relations as “is a”, “kind of”, “part of”, “subclass of”, and so on; and non-hierarchical relations as “affects”, “has function”, “instrument”, “purpose”, etc. Information about “descendants” and “ancestors” is also included in this table, so that the user gets a general idea of the position of the searched concept in the ontology (cf. Figure 33).

Finally, and as shown in Figure 34, linguistic and terminological information is added to the concept in English and Spanish: the “term” in each language, “term type” data, “part of speech”, “number”,

⁵⁹ <http://www.ttt.org/clsframe/index.html>

⁶⁰ <http://www.ttt.org/clsframe/reltef.html>

“definition”, “context” of use, and administrative data, “origination date” and “originator”. Apart from this linguistic data, a “reliability code” that goes from 0 (no reliability at all) until 10 (highly reliable) is introduced by the authors in order to inform the user of the authoritative information and the evaluation carried out.

[ADULT-T-CELL-LEUKEMIA-LYMPHOMA](#)

[ADVANCED-CANCER](#)

[AGGRESSIVE-NON-HODGKINS-LYMPHOMA](#)

[AGRANULOCYTE](#)

[AIDS](#)

[AIR-CONTRAST-X-RAY](#)

[ALBINISM](#)

[ALCOHOLIC-BEVERAGE](#)

[ALEUKEMIA](#)

[ALKALINE-PHOSPHATASE-TEST](#)

[ALKALINE-PHOSPHATASE](#)

[ALKYLATING-DRUG](#)

[ALLOGENEIC-BONE-MARROW-TRANSPLANT](#)

[ALPHA-FETOPROTEIN-TEST](#)

[ALPHA-INTERFERON](#)

[ALTERNATIVE-TREATMENT](#)

[ALTRETAMINE](#)

[ALVEOLAR-CELL-LUNG-CANCER](#)

[ALVEOLAR-SOFT-PART-SARCOMA](#)

[ALVEOLUS](#)

[AMSACRINE](#)

[AMYLOIDOSIS](#)

[ANAL-CANCER](#)

[ANALGESIC-PUMP](#)

[ANAPLASIA](#)

[ANAPLASTIC-ASTROCYTOMA](#)

[ANAPLASTIC-OLIGODENDROGLIOMA](#)

[ANAPLASTIC-THYROID-CARCINOMA](#)

[ANASTOMOSIS](#)

[ANDROGEN](#)

[ANEMIA](#)

[ANESTHESIA](#)

[ANGIOGRAM](#)

[ANGIOGRAPHY](#)

[ANGIOSARCOMA](#)

[ANOSCOPY](#)

[ANTERIOR-EXENTERATION](#)

[ANTERIOR-RHIZOTOMY](#)

[ANTIANDROGEN](#)

[ANTIBODY](#)

[ANTIGEN](#)

[ANUS](#)

[APLASIA](#)

[APLASTIC-ANEMIA](#)

[APPENDECTOMY](#)

[APPENDIX](#)

[ARC-RADIATION-THERAPY](#)

AIR-CONTRAST-X-RAY

check date: 12/08/2001

checker: Pamela

subject field: medicine: diagnostic procedure

origination date: 16/01/2001

originator: Claudia

Conceptual Structures	
ISA	CONTRAST-X-RAY
SUBCLASSES	PNEUMOALVEOLOGY PNEUMOANGIOGRAPHY
DESCENDANTS	PNEUMOALVEOLOGY PNEUMOANGIOGRAPHY
ANCESTORS	ALL CONTRAST-X-RAY DIAGNOSTIC-PROCEDURE EVENT HEALTH-SERVE INVASIVE-DIAGNOSTIC-TEST MEDICAL-SERVE PROFESSIONAL-SERVICE-EVENT SERVICE-EVENT SOCIAL-EVENT WORK-ACTIVITY X-RAY
RELATIONS	HAS-PARTS : INSUFFLATION USES : CONTRAST-MEDIUM

Figure 33: Results from the search of “air-contrast-x-ray” in the OncoTerm resource

English	
air-contrast X rays	<p>term type: main entry term</p> <p>part of speech: noun</p> <p>number: singular</p> <p>reliability code: 10</p> <p>definition: X ray that uses inflation with air to obtain a better outline of some parts of the body. The air can be inhaled, swallowed, injected, or ingested by drinking a carbonated beverage, after which the X ray is taken. This procedure may be used to view the bladder, stomach, brain, spinal column, and mediastinum. (<i>en</i>)</p>
Spanish	
neumografia	<p>term type: main entry term</p> <p>part of speech: noun</p> <p>number: singular</p> <p>gender: feminine</p> <p>reliability code: 10</p> <p>definition: radiografía en la que se utiliza el aire como medio de contraste para destacar determinadas partes del cuerpo. El aire puede ser inhalado, succionado, inyectado o ingerido por medio de una bebida carbonada. Este procedimiento se emplea para visualizar la vejiga, el estómago el cerebro, la columna vertebral y el mediastino. (<i>es</i>)</p>
	<p>term type: synonym</p> <p>part of speech: noun</p>

Figure 34: Linguistic dimension of the OncoTerm resource

Systems of representation of multilingual information. The OncoTerm terminological database follows the term base data model of the CLS Framework, as mentioned before, and the relational database manager Reltef™.

The CLS Framework was designed in order to deal with the structure and content of terminological databases. It is a logical organization of the ISO 12620 data categories, i.e. it takes some data categories identified in the ISO 12620 document which are relevant for representing terminological information and arranges them according to the needs of such a resource. This data category selection has resulted in the development of the MARTIF standard (ISO 12200:1999) that, in turn, it enables the exchange of data among terminological resources. The CLS Framework can be used for representing the information of the existing term bases, designing new ones, and sharing terminological data. The CLS Framework includes the application Reltef™, a model consisting of an Entity Relation diagram and a set of tables and relationships, which is in charge of the data recovery and maintenance of the database.

By using the OntoTerm™ tool of the CLS Framework and the Reltef™ terminology database manager, the concepts of the oncology domain are arranged in an ontology in the Ontology Editor, and linked to the terminological information stored in the TermBase Editor.

Evaluation methods. Experts, terminologists and translators of the different Universities involved in this research are responsible for the evaluation of the ontology and the terminological data base, which was carried out manually.

URL: <http://www.ugr.es/~oncoterm/>

Accessed February 2007.

11. Conclusions and Summarizing Tables

11.1 Main Conclusions to the Multilingual Resources Survey

In the first part of this Deliverable 2.4.1, we have analyzed some Localization Approaches of Lexical Resources (LRs) and Ontologies in order to obtain an overview of current Localization methodologies. The conclusions we can extract have been summarized in two sections: conclusions from the Localization Approaches of LRs; and conclusions from the Localization Approaches of Ontologies.

11.1.1 Localization Approaches of LRs

LRs analyzed in this survey were: FAOTERM, FishBase, Eurodicautom, AGROVOC, and Eurovoc, and the main conclusions have been listed below:

1. Some of the strategies, techniques and tools used for the localization process of LRs could be reused in certain stages of the localization process of ontologies.
2. The use of available lexical resources relevant for the domain ontology, as well as text repositories is also crucial in the ontology localization process. In the same way, most of the translation supporting tools, editors and workflow automation applications used for the localization process of these LRs could be adapted and reused for ontologies, since most of those applications are independent of domain and language.
3. Regarding the search options of the interfaces, all of them can be adapted for an ontological resource, and it only depends on the quantity of linguistic information that is to be included in the ontology. It is recommendable to determine the amount of linguistic information in the first stages of the ontology development in order to extract the pertinent knowledge from the available LRs and text repositories, counting on the various supporting tools.
4. As for the E/R schema, the representation diagram identified for FAOTERM or AGROVOC (Figure 15) could also be adopted for the representation of multilingual information in ontologies. Should this be the case, the ontology would be provided with multilingual information associated with ontology elements, and the linguistic information would be stored in the ontology model.

11.1.2 Localization Approaches of Ontologies

In this section, we summarize the conclusions drawn from the analysis of the ontological resources, methodologies and tools included in this deliverable, i.e., EuroWordNet, Termontography, LabelTranslator, OntoLing, GENOMA-KB, and OncoTerm. According to them, three different localization approaches are proposed depending on the purpose of the resource and the development stage in which multilinguality is provided to the ontology.

- a) Localization approach based on monolingual ontologies linked to each other
- b) Localization approach based on the *in situ* translation process of monolingual ontologies
- c) Localization approach based on a language independent ontology linked to a linguistic model

a) Localization approach based on monolingual ontologies linked to each other

The localization approach represented by EuroWordNet (section 9.1) has proven to be the most recommendable if the purpose of the ontology is to organize or model general knowledge. The decision of following this approach has to be taken during the first stages of the ontology development. In this approach, two different types of ontologies will be needed. On the one hand, a language neutral conceptualization that will enable links or mappings among all monolingual ontologies. On the other hand, various language dependent conceptualizations because of its capability to capture language and culture specificities. The Entity/Relation schema proposed for this localization approach is the one identified in Figure 23 that conforms to the EWN representation system of multilingual information.

b) Localization approach based on the *in situ* translation process of monolingual ontologies

The main advantage of this approach represented by the LabelTranslator (section 10.2) and OntoLing (section 10.3) tools is that it provides multilinguality to existing ontologies in one natural language. The purpose of the ontology is not relevant, as long as general and specific multilingual lexical resources are available for the localizing task. The ontology taken as starting point will normally be a language dependent conceptualization. This fact will require the use of mechanisms to tackle the problem of disparities among conceptualizations in different languages. This could be solved by the addition of comments or notes in natural language, or at a conceptual level, by the introduction, for example, of language specific modules in the ontology. This second option would need further research. The representation schema of multilingual information will depend on the representation schema of the original ontology (or the ontology that undergoes the localization process), and so multilingual data will be added to the linguistic data already available in the ontology. It is highly probable that the representation schema adopted when following this approach is the one that corresponds to FAOTERM or AGROVOC Figure 15, because most existing monolingual ontologies present linguistic information embedded in the ontology, so that, in consequence, multilingual information will also be placed inside the ontology.

c) Localization approach based on a language independent ontology linked to a linguistic model

Regarding the third localization approach identified in this research in the GENOMA-KB (section 10.4), and OncoTerm (10.5) resources, it is based on a language independent conceptualization – in which no linguistic information is contained- linked to a resource whose function is to provide the referred conceptualization with multilingual data. The quantity of linguistic data to be included in that resource will be determined by the purpose and linguistic needs of the resource. Such an approach is preferable when modelling highly specific domains of knowledge, in which one conceptualization is sharable among several languages. This would be the case whenever one and the same conceptualization fits in different knowledge structures represented by several language and cultural perspectives. This approach can only be taken into account if the ontology is being developed from the start and multilinguality is included at the same time, as suggested by the Termontography methodology (section 10.1). The most appropriate representation schema is the one introduced by GENOMA-KB and OncoTerm (Figure 30). In this way, linguistic data are kept out of the ontology thus simplifying the addition of as much linguistic information as needed, or the addition of an entire new language.

All approaches presented here would benefit from the **localization of the interface messages**, i.e. by including the option of changing the messages of the interface to all languages in which the content of the resource is available. As a result, both lexical and ontological resources become more user friendly, thus widening the range of users.

A summary of the main information related to the surveyed LRs and Ontological resources has been included in Table 11, Table 12, and Table 13 in this section. The purpose of these tables is to offer a quick overview of the set of resources regarding administrative data as developers name or

number of contained records; aims, languages and domains involved in the localization process; as well as steps and tools employed for the localizing task.

11.2 Summarizing tables

Table 11: General description of approaches

Feature	FAOTERM	FishBase	Eurodicautom	AGROVOC	EUROVOC	EWN	Termonotography	LabelTranslator	OntoLing	GENOMA-KB	OncoTerm
Developer	FAO	Pauly & Froese	EC	FAO	EU	Universities of Holland, Spain, Italy, England, France, Germany, Czech Republic, Estonia	CVC at the Erasmushogeschool Brussel within the FF POIROT project (IST 2001-38248)	Ontology Engineering Group at the UPM, within the Esperanto project (IST-2001-34373)	AI Research Group, Department of Computer Science, Systems and Production of the University of Rome, Tor Vergata	IULATerm at the Institute of Applied Linguistics of the UPF	Universities of Granada, Malaga and Valladolid. Hospital Virgen de las Nieves in Granada
Launching date or project conclusion date (pcd)	2001 on the Internet	1988	1973	1982	1984	1999	2004 (pcd)	2005 (pcd)	2006	2003 (pcd)	2002
Current records	70,000	Aprox.30,000 species, 222,000 common names, 43,000 pictures, 39,000 references,	5,5 Mio entries	Aprox. 300,000	Aprox. 7,000 descriptors	Aprox. 90,000 per language	Depends on ontology labels	Depends on ontology labels	Depends on ontology labels	Depends on ontology labels	1896 concepts and 4000 terms

Table 12: Aims, languages and domains involved in the resources

Feature	FAOTERM	FishBase	Eurodicautom	AGROVOC	EUROVOC	EWN	Termonotography	LabelTranslator	OntoLing	GENOMA-KB	OncoTerm
Aims	Communication & public information	Unify terminology	Solve EC translators' terminological needs	Standardize indexing process	Standardize indexing process & solve EC translators' terminological needs	Improve multilingual queries	Knowledge management & representation	Translation of ontology labels	Translation of ontology labels, and inclusion of definition and synonyms in natural lang.	Information retrieval & terminology control	Information retrieval & terminology control
Languages involved	en, fr, es, it, ar, zh	en, es, fr, de, it, pt, nl, el, sv, zh, ru, vi, th, ms/id	da, fi, el, pt, nl, fr, it, es, en, de, la, sv	en, fr, es, ar, pt, zh, th, cs, sk, ja	es, cs, da, de, el, en, fr, it, lv, hu, nl, pl, pt, sl, fi, sv	en, nl, it, es, fr, de, cs, et	en, nl, fr, it	es, en, de	en, de, es, fr, da, it, hu, ru, sv	es, en, cat	es, en
Domains involved	Food, agriculture, forestry, fisheries	Ichthyology and fisheries	Human knowledge & 48 subject fields related to EU policy	Agriculture, forestry, fisheries, nutrition, food, environment...	All fields of the EU (main ones: law & EU legislation)	General purpose lexicon	Depends on ontology domain	Depends on ontology domain	Depends on ontology domain	Human genoma	Oncology

Table 13: Steps and tools used for localization

Feature	FAOTERM	FishBase	Eurodicautom	AGROVOC	EUROVOC	EWN	Termonotography	LabelTranslator	OntoLing	GENOMA-KB	OncoTerm
Source language for the creation of the resource	en	en	en, fr	en	fr, en (in the recent years)	en, nl, it, es, fr, de, cs, et	All languages involved in the process	Ontology source language	Ontology source language	None (language independent conceptualization)	None (language independent conceptualization)
Translation process	semi-automatic	manually (fixed text); automatic (free text)	manually (initially); semi-automatic (in recent years)	manually	semi-automatic	manually & semi-automatic	semi-automatic	semi-automatic	semi-automatic	semi-automatic	semi-automatic
LRs used for the translation task	international multilingual databases	multilingual glossaries; multilingual text repositories	multilingual text repositories; multilingual glossaries, dictionaries and thesauri	multilingual thesauri, lexicons, dictionaries & encyclopedias	multilingual text repositories; multilingual encyclopedias, dictionaries, term banks & glossaries	monolingual & bilingual dictionaries; taxonomies; databases	not defined	EWN; Wikipedia; Babelfish	WORDNET DICT	monolingual & bilingual dictionaries; multilingual text repositories, etc	monolingual & bilingual dictionaries; multilingual text repositories, etc
Translation supporting tools	translation memories; text alignment tools; term extraction tools, glossary building & maintaining tool, editors	machine translation	translation memories; linguistic data processor; machine translation; workflow automation	not defined	translation memories; machine translation; data processors; voice recognition tools; editors; searcher and concordance tools	term mappers	web crawler; keyword extractors; automatic aligner; term identifier; translation extractors	It is in itself a translation supporting tool	It is in itself a translation supporting tool	term mappers; term extractors, workflow automation	concordancer, term extractors, workflow automation

Representation of multilingual information	available	not available	not available	available	not available	available	not available	depends on tool being localized	depends on tool being localized	available	available
Evaluation method	manual	manual	manual	semi-automatic	manual	manual	manual	manual	manual	manual	manual

12. Representation of multilinguality in NeOn

12.1 Introduction

Every localization process has as a result the creation of a multilingual resource, such as the ones we have analysed in sections 5 to 11. **We contend that a resource is multilingual when it is available in more than one natural language.** This document proposes a meta-model for the representation of multilinguality and associated linguistic information in NeOn. As stated in **D1.1.1 Networked Ontology Model – Draft**, *Metamodels are used for the specification of modelling languages in a standardized, platform independent manner. (...) The term meta-model is chosen, as a meta-model refers to model of a language, whereas the instances of the meta-model are referred to as models.* In our case, models thus refer to the actual ontologies. In this document we aim at illustrating multilingual ontology meta-models, accompanied by an example of the corresponding ontology models, taking as the grounding the ontology meta-model defined in WP1.

Within ontology architecture, multilinguality occurs at different levels of the information structure, namely:

- 1) Interface level
- 2) Metadata level
- 3) Knowledge Representation level
- 4) Data level⁶¹

For our modelling purposes we will stay within the confinements of an ontology, and concentrate on the first three levels.

In addition to the linguistic and terminological knowledge representation standards described in chapter 2, we begin this survey by introducing the existing localization standards we will take into account when encoding multilinguality, in order to guarantee, in the first place, interoperability (section 12.2). Then, we present a list of requirements expressed in the different WPs in NeOn that need to be considered for representing multilinguality in the desired way (section 12.3). In the following section 12.4 we justify our election of a three layered approach and present the evaluation criteria we have followed in order to come up with a definitive proposal of multilingual information representation in NeOn. In the following chapters (13-15) proposals of representation of multilinguality at Interface, OMV and Knowledge Representation levels are described in detail, together with a discussion of their advantages and disadvantages. At the end of section 15 we have included a table summarizing the criteria that determine the advantages and limitations of the possible Multilingual Ontology Meta-models. Finally, in section 16 we present the Multilingual Ontology Meta-model agreed for representing multilinguality in NeOn. However, since this model will not be implemented by month 18 (August 2007) because of time constraints, a first prototype has been proposed to support multilinguality in the current version of the NeOn toolkit. Architecture and functionalities of this first prototype have been described in section 17.

12.2 Standardization of localization

As mentioned above, localization has recently deserved the ontology community's attention in that, once the ontologies are created in a natural language they are often localized into another natural language to make them accessible to different communities. For this particular purpose and the general purpose of interoperability,

⁶¹ At this stage of the document, the data level has not been included.

there exist a number of standards in various stages of development. We present here some standards that can be taken into consideration for NeOn purposes.

12.2.1 TMX (Translation Memory Exchange)

TMX is an XML-compliant standard method for the description of translation memory data that is being exchanged among translation tools. TMX has been developed by OSCAR (Open Standards for Container/Content Allowing Re-use), a LISA (Localization Industry Standards Association) Special Interest Group.

TMX files are always in Unicode. They can use one of the three encoding methods: UTF-16, UTF-8 or ISO-646.

General structure of a TMX document:

A TMX document is enclosed in a <tmx> root element. The <tmx> element contains two elements: <header> and <body>.

Structural elements:

The **<header>** contains information (meta-data) about the TMX document.

It may contain one or more <note> (*note* element for comments), <ude> (*user-defined encoding* element for specifying user-defined characters) or <prop> (*property* element for defining properties of parent elements) elements.

A complete description of all Attributes is to be found in Table 14.

Table 14: Compulsory and Optional Attributes of the <header>

Attributes	Optional Attributes
creationtool	o-encoding (original or preferred code set of the data)
creationtoolversion	creationdate
segtype	creationid
o-tmf (<i>original translation memory format</i>)	changedate
adminlang	changeid
srclang (source language)	
datatype	

This has been illustrated in the following example of a <header>:

```
<header
creationtool="Transit"
creationtoolversion="3.0"
datatype="Transit"
segtype="block"
adminlang="en"
srclang="en-gb"
o-tmf="Transit"
creationdate="20010507T083458Z"
```

```

creationid="XTRA-BI"
o-encoding="Unicode"
>
<prop type="Project">Traduccion de prueba</prop>
</header>

```

The **<body>** contains the set of **<tu>** (*translation unit*) elements. Each **<tu>** contains the **<tuv>** (*translation unit variant*) element, i.e. the information in one of the languages of the resource. The text itself is stored in the **<seg>** (*segment*) element. As in the case of the **<header>** element, the **<note>** and **<prop>** elements are used in order to include information specific to each **<tuv>**.

Example of a **<tu>** in a **<body>**:

```

<tu>
<tuv xml:lang="es">
  <seg>Hola, mundo.</seg> </tuv>
<tuv xml:lang="en">
  <seg>Hello, world.</seg> </tuv>
<tuv xml:lang="eu">
  <seg>Kaixo, mundua.</seg> </tuv>
</tu>

```

The **<map/>** element is used to specify user-defined characters and its properties. Required attribute is: Unicode

The **<note>** element allows for comments and contains text.

The **<prop>** element defines properties of parent elements. Required attribute is type.

The **<seg>** element contains the text data. A segment can contain markup content elements: The **<bpt>**, **<ept>**, **<it>**, and **<ph>** elements allow you to encapsulate original native inline codes. The **<hi>** element allows you to add extra markup not related to existing inline codes. And the **<sub>** element, used inside encapsulated inline code, allows you to delimits embedded text. These elements are considered "inline elements" because they appear inside a segment.

Table 15: Inline Elements

Inline elements

<bpt>	<i>begin paired tag</i> , used to delimit the beginning of a paired sequence of native codes.
<ept>	<i>end paired tag</i> , used to delimit the end of a paired sequence of native codes.
<hi>	<i>highlight</i> , used to delimit a text with special meaning, as for example, a proper name.
<it>	<i>isolated tag</i> , used to delimit a beginning/ending sequence of native codes that does not have its corresponding ending/beginning within the segment.
<ph>	<i>placeholder</i> , used to delimit a sequence of native standalone codes in the segment.
<sub>	<i>sub-flow</i> , used to delimit sub-flow text inside a sequence of native code.
<ut>	<i>unknown tag</i> , used to delimit a sequence of native unknown codes in the segment.

Table 16: TMX Attributes

Attributes

adminlang (<i>Administrative language</i>)	default language for the administrative data
---	--

assoc (Association)	association of a <ph> with the text prior or after
changedate	date in ISO format of modification of the element
code	code-point value corresponding to the Unicode character of a given <map/> element
creationdate	creation date in ISO format of the element
creationid	the id of the user who created the element
creationtool	tool that created the TMX document
creationtoolversion	version of the tool
datatype	type of data contained in the element (e.g. "unknown", "html", "java", "plaintext", etc.)
ent (Entity)	entity name of the character defined by a <map/> element
i (Internal matching)	used to pair <bpt> with <ept> elements
lastusedate	when the content of a <tu> or <tuv> element was used for the last time
name	name of a <ude> element.
o-encoding (Original encoding)	original code of the text before Unicode
o-tmf (Original translation memory format)	format of the translation memory file
pos (Position)	indicates whether a <it> is a beginning or an ending tag
segtype (Segment type)	kind of segmentation used in a <tu> element. Values are: "block", "paragraph", "sentence", or "phrase".
srclang (Source language)	language of the source text
subst (Substitution text)	alternative string for the character defined in a given <map/> element
tuid (Translation unit identifier)	identifier for the <tu> element
type	kind of data of a <prop>, <bpt>, <ph>, <hi>, <sub>, or <it>. Possible values are: "bold", "time", "fnote" (<i>Footnote</i>), etc.
unicode	Unicode character value o a <map/>
usagecount	number of times a <tu> or <tuv> content have been accessed in the TM
version (TMX version)	version of the TMX format
x (External matching)	matches inline elements between each <tuv> of a <tu>
xml:lang (Language)	language of the text of a given element

Relevant bibliographic references: <http://www.lisa.org/standards/tmx/>

12.2.2 XLIFF

XLIFF, which stands for XML Localization Interchange File Format, is a format for exchanging localization data between companies, such as a software publisher and a localization vendor, or between localization tools, such as translation memory (TM) systems and machine translation (MT) systems.

XLIFF is an XML-based format that enables translators to concentrate on the text to be translated. Likewise, since it's a standard, manipulating XLIFF files makes localization engineering easier: once you have converters written for your source file formats, you can simply write new tools to deal with XLIFF and not worry about the original file format. It also supports a full localization process by providing tags and attributes for review comments, the translation status of individual strings, and metrics such as word counts of the source sentences.

The XLIFF format grew out of a collaboration between a number of companies, including Sun Microsystems, but was soon brought under the management of an [OASIS Technical Committee](#). In April 2002, the first Committee Specification for XLIFF

was published. This is available at <http://www.oasis-open.org/committees/xliff/documents/xliff-specification.htm>.

The XLIFF format aims to:

- Separate localizable text from formatting.
- Enable multiple tools to work on source strings and add to the data about the string.
- Store information that is helpful in supporting a localization process.

The XLIFF File

In its most basic form, the XLIFF file consists of one or more file elements. Each of these contains a header and a body section. The header contains project data, such as contact information, project phases, pointers to reference material, and information on the skeleton file (explained below). The body section contains `trans-unit` elements--the main elements in an XLIFF file.

The `trans-unit` elements store localizable text and its translations. These elements represent segments (usually sentences in the source file that can be translated reasonably independently). The `trans-unit` elements contain `source`, `target`, `alt-trans`, and a handful of other elements. The example below shows how they would be used.

Example of a `trans-unit` Element

...

```
<trans-unit id="n1">
<source>This is a sentence.</source>
<target xml:lang="fr">Translation of "This is a sentence."</target>
<alt-trans match-quality="100%" tool="TM_System">
  <source>This is a sentence.</source>
  <target xml:lang="fr">TM match for "This is a sentence."</target>
</alt-trans>
<alt-trans match-quality="70%" tool="TM_System">
  <source>This is a short sentence.</source>
  <target xml:lang="fr">Fuzzy TM match for "This is a sentence."</target>
</alt-trans>
</trans-unit>
```

...

This example shows a pseudo-translated segment. The `trans-unit` element contains an `id` attribute used to determine where the segment goes in the original document. The `trans-unit` element has a `source` and a `target` element as children. The `source` element represents the source text (the text to be translated) in the original document. The `target` element represents the currently accepted translation of the source after linguistic review has taken place.

The example also shows the `alt-trans` elements. These represent translation alternatives for the `source` segment in the `trans-unit` element. A translation alternative is a translation found in a translation memory, a translation generated by a

machine translation system, or a translation suggested by a translator or reviewer. These elements contain `source` and `target` elements. In this example, `target` elements are the suggested translations of the `trans-unit source`. The `source` element represents the text that was matched against, from a TM system, for example.

The `alt-trans` element contains attributes such as `match-quality` and `tool`. These provide information about the alternative translations, such as which tool produced them, or in the case of `match-quality`, a measure of the quality of the translation. The algorithm for generating the `match-quality` value in a given `alt-trans` element is specific to the tool that generated it. However, for a translation memory system, it is typically the percentage of words in the `source` element that match the `source` from its database

Relevant bibliographic references:

http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff

12.2.3 MLIF

MLIF is a proposed ISO standard for the representation of multilingual information. The main objective of MLIF is to provide “a common conceptual model and a platform allowing interoperability among several translation and localization standards (...)” (Cruz-Lara *et al.* 2006). As with other standards, for example TMF (outlined in section 2.1 of this Deliverable), MLIF introduces a meta-model, which, in combination with some data categories from ISO 12620:1999, allows interoperability and exchange with multilingual applications and corpora. MLIF can also be linked to other standards, if required by the domain, as the MAF, standard for morphological description (ISO CD 24611), SYnaf, for syntactical annotation (ISO WD 24615) or TMF, for terminological description (ISO 16642:2003). MLIF is also able to interoperate with other existing standards, as TMX (Translation Memory eXchange) and XLIFF (XML Localisation Interchange File Format), and it is therefore considered as a parent format or common framework for all of them.

MLIF Meta-model

The MLIF Meta-model consists of the following components: a Multilingual Data Collection (MLDC) that contains a collection of MultiLingual Components (MULTI), and is linked to a Global Information (GI) component that contains data related to technical and administrative information.

The MULTI component represents a unique multilingual entry, and several MonoLingual Components (MONO), each containing information related to one language.

The Segmentation Component (SEG) allows for any level of segmentation of the textual information.

Finally, the History Component (HISTORY) can be anchored to several components, as can be seen in Figure 35, and contains information about the creation, modification, etc., of a specific component, as well as data related to the author and date.

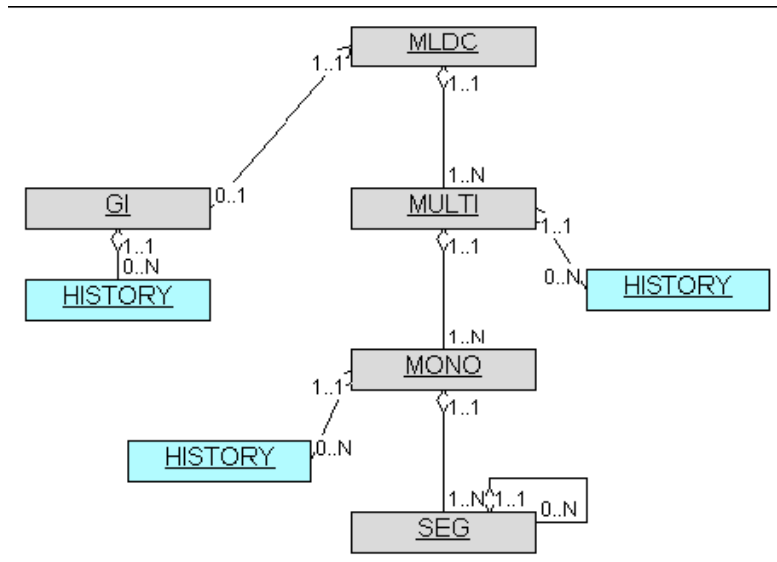


Figure 35: MLIF Metamodel

Possible Data Categories for MLIF

Since MLIF aims at providing a generic structure for other models, the elements or attributes proposed in Figure 36 will be explicitly defined or not, depending on the domain. Then, the different models will define their own elements making use of the extensibility criterion of this meta-model.

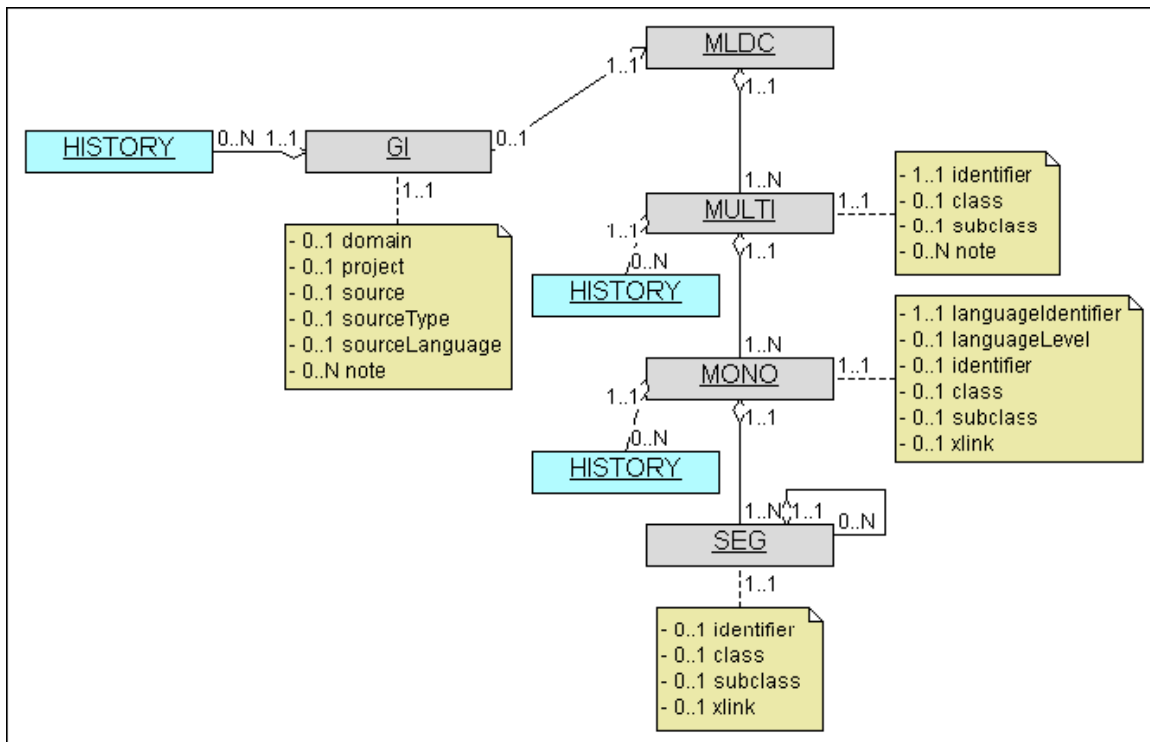


Figure 36: MLIF Metamodel with related Data Categories

Finally, we have introduced a simple example of the MLIF in XML

```
<?xml version="1.0" encoding="UTF-8"?>
<MLDC xmlns:xlink="http://www.w3.org/1999/xlink">
  <GI>
    <domain>multilingual</domain>
    <project>mlif</project>
    <sourceLanguage>fr</sourceLanguage>
    <source>source</source>
    <HISTORY>
      <transaction>origination</transaction>
      <author>Jonathan VEITMANN</author>
      <date>20070101T163812Z</date>
    </HISTORY>
  </GI>
  <MULTI xml:id="1" class="File">
    <HISTORY>
      <transaction>origination</transaction>
      <author>Samuel CRUZ-LARA</author>
      <date>20070101T191612Z</date>
    </HISTORY>
    <MONO xml:lang="en" xml:id="en_all_1" languageLevel="all">
      <SEG class="category" subclass="New Work Item Proposal">
        New Work Item Proposal</SEG>
      <SEG class="titre">
        <SEG xlink:href="/mod_file/upload/MLIF-draft.pdf">Draft</SEG>
      </SEG>
    </MONO>
  </MULTI>
  <MULTI xml:id="15" class="Lien">
    <HISTORY>
      <transaction>origination</transaction>
      <author>Julien DUCRET</author>
      <date>20070101T163812Z</date>
    </HISTORY>
    <MONO xml:lang="en" languageLevel="all" xml:id="en_all_15">
      <SEG class="category" subclass="Application">Application</SEG>
      <SEG class="titre">MLIF editor</SEG>
      <SEG class="primaryText">
        java tools which allows to create and edit mlif file</SEG>
      <SEG xlink:href="/java/mlif/index.php">Link</SEG>
    </MONO>
    <MONO xml:lang="fr" languageLevel="all" xml:id="fr_all_15"
    xlink:label="source" xlink:href="#en_all_15">
      <SEG class="category" subclass="Application">Application</SEG>
```

```

<SEG class="titre">MLIF éditeur</SEG>
<SEG class="primaryText"> application java qui vous permet de créer et
    modifier un fichier MLIF</SEG>
<SEG xlink:href="/java/mlif/index.php">Lien</SEG>
</MONO>
</MULTI>

```

Relevant bibliographic references:

ISO/TC37/SC4 internal document

Cruz-Lara, S., N. Bellalem, J. Ducret and I. Kramer. (2006). "Standardizing the management and the representation of multilingual data: the MultiLingual Information Framework". International Workshop on Language Resources for Translation work, Research and Training, Genoa, Italy, pp. 35-38. On-line available at: <http://hal.inria.fr/inria-00105653/en/>

Cruz-Lara, S., N. Bellalem, J. Ducret & I. Kramer. (2006). "Interoperability between translation memories and localization tools by using the MultiLingual Information Framework" In European Association for Machine Translation - EAMT 2006, Oslo/Norway
On-line available at: <http://www.mt-archive.info/EAMT-2006-Cruz-Lara.pdf>

12.3 Requirements for multilinguality in NeOn and restrictions

In this section we evaluate the requirements for multilingual information representation included in the different NeOn WPs, in order to extract the main restrictions we have to consider before coming up with a final proposal.

From each requirement we have derived the main restrictions for the proposal. Requirements are written in italics to differentiate them from our evaluation.

WP1

- WP1 (D1.1.1)
 - *With respect to the case studies, modularizing the fishery ontology into regional or economic regional differences is probably useful. Also, ideally, to separate (or modularize) the multilingual side of ontology would also be promising. For example, one multilingual ontology, where concepts have names (attributes) in several languages (multilingual layers). **The structure of the ontology is meant to be independent of the multilingual layer (which means it has been modularized), possibly with a strict 1-1 correspondence between the various languages and single ownership of the entire ontology (...).***
 - *A particular use case in the AGROVOC thesaurus (to be converted into an ontology). AGROVOC (...) does not grow simultaneously in all languages. Given its collaborative nature, it needs a modularization tool (in terms both of structure and languages layers, to be able to have different **authors and/or institutions work on it**) that is built on the **modularization model**.*

In WP1, it is understood that the structure of the ontology will exist independently from a multilingual layer. For certain resources, as in the case of the AGROVOC thesaurus

from FAO, it is also understood that the so-called multilingual layer could be managed by different authors and institutions in different locations. It has to be borne in mind that those “authors” may or *may not* know how to manage or deal with ontologies.

- 5.1.1 *What do mappings define?*
 - (...) we see mappings as axioms that define a **semantic relation** between elements in different ontologies. A number of different kinds of semantic relations have been proposed. Most common are the following kinds of semantic relations:
 - Equivalence (...)
 - Containment (...)
 - Overlap (...)
- Adding these negative versions of the relations leaves us with eight semantic relations that **cover all existing proposals for mapping languages**.*

According to this requirement, no “translate” semantic relation exists, which means that there would be no way to distinguish different types of “equivalence” relation mappings. Therefore, this leaves no margin for a multilingual representation system based on mappings to link ontologies in different languages.

WP6

- WP6 (D6.1.1):
 - 3.2.2.5 Multilinguality support
 - *A fundamental feature, shed by the NeOn pharmaceutical and fishery case studies, is support for multilingual ontologies. It is key for their respective domains that the models described by these ontologies are available in a series of different languages. Possible ways to implement multilinguality is by means of **contextualized ontologies**.*
- Requirement 3.2.15
 - *NeOn shall support multilingual ontologies, **implement multilinguality by means of contextualized ontologies**.*

Following the requirements in WP6, multilinguality is implemented “by means of contextualized ontologies”. According to the definition of contextualized ontology given by WP3, “A contextual ontology is a pair of OWL ontologies, a set of context mappings”, the proposal of an ontology meta-model would depend on a mappings meta-model. However, as already stated in WP1, the set of mappings identified for NeOn does not include mappings with the semantic relation “translate”, and it is not viable to use mappings to represent multilinguality.

WP7

- WP7 (D7.1.1):
 - 4.4.7 Multilinguality

- *Editors also deal with the multilingual aspect of FAO resources. They should be able to:*
 - **incorporate a new language** for an entire ontology;
 - *select at least two languages (or more, if required), one in view mode, the other in editing mode;*
 - **add/edit/delete multilingual labels** to individual concepts;
 - **cope with specificities of translation** (i.e., no lexicalization available for concepts, available lexicalization corresponds to more than once concept or conversely, several lexicalizations are possible).

The requirement of including a **new language** for an entire ontology also implies that the **complexity** associated to it should be taken into consideration when proposing a Multilingual Ontology Meta-model (MOM).

- Use case 4.6.15
 - **Add a new language to multilingual ontologies:** (...) user is requested to specify which elements of the ontology should be multilingual: user can select **classes, instances, properties, or the entire ontology.**

This requirement implies that not only concept labels, but also attributes, relations, axioms, etc. have to be able to support multilinguality. In this sense, we must bear in mind that the ontology meta-model will have to cope with a considerably high number of ontology components.

When adding a new language to the already multilingual ontology, we may also want to reflect this information at the metadata of such a representation. This implies a modification of the ontology metadata in the Ontology Metadata Vocabulary (OMV) (Hartman and Palma 2006) in order to report about the different ontology elements that are have natural language information associated to them.

- Use case 5.3.8
 - *Change language of the interface*

In order to be able to change the language of the interface, messages of the user interface have to be multilingual. In this case we have to consider which way is the most practical to visualize the information, and which modifications have to be carried out in the visualizing code in order to add a new language.

- Use case 5.3.9
 - *Change language of the resource shown*

The latter requirement implies that the information in the so called “resource” has to be presented in different languages. There are two main options to meet this requirement: either the information of the Knowledge Base (KB) already exists in different languages, or there is a monolingual KB and a lexical resource that enables a translation process.

- AGROVOC and the OWL Web Ontology Language. The Agriculture Ontology Service / Concept Server OWL model (document from WP7)

- The multilingual issue
 - *To prepare AGROVOC for use as an ontology, it is essential to represent concepts by minimizing bias towards a given language or family of languages. That is, to the extent possible, **meaning is considered independently of its realization in a particular language.** Each language would then be able to express the domain concepts for which it had lexicalizations and for which others may not. A terminology that simply translated the terms in a given language, such as English, would miss out on concepts that were not lexicalized in that language.*
- The basic model
 - (...) *lexicalizations of the concepts will occur as instances of the class `c_lexicalization`.*
- The lexicalizations
 - (...) *we opt for organizing the class lexicalization into subclasses by language, while keep using the `rdfs:label` to mark instances by language:*
 - `---c:lexicalization`
 - `---c:lexicalization_EN`
 - `---c:lexicalization_ES`
 - `---c:lexicalization_CZ`
- Lexicalizations of properties
 - (...) *we lexicalize properties with one label respectively for each language.*
- Definitions
 - *If a definition is available in several languages at one source, there will be an `rdfs:label` for each language of the definition.*
- Disambiguation⁶²
 - (...) *domain specific sub-relationships should help on that. For the purpose of indexing, definitions, scope notes, comments and relations all contribute to make clear what term to use (...).*

We have considered these requirements for multilinguality identified in the “AGROVOC and the OWL Web Ontology Language”⁶³ document as hints of possible requirements from WP7 to NeOn. The main conclusion extracted from their evaluation is that the ontology should be language independent, that multilingual definitions may accompany labels of concepts, and that “scope notes”, “comments” and “relations” should help identify specific concepts in an ontology. The option of having one ontology in each language and mappings between them has to be consequently discarded.

WP8

- WP8 (D8.1.1):
 - 5.3.4.3 Multilinguality and ontology localization in Semantic Nomenclator:
 - *Reference ontology should admit different official languages in Spain (Spanish, Catalan, Galician and Basque). **When the reference ontology is developed, NeOn should suggest candidates for the ontology label ...***

⁶² Some of the sections in this point are not included in the document “AGROVOC and the OWL”, but resulted from specific questions on this document to our FAO partners.

⁶³ <http://www.neon-project.org/ACollab/drafting/index.php?id=81>

According to this requirement from WP8, one can deduce that the interface content is multilingual, and that the resource is provided multilingually during the design time. As in the first requirement of WP7, that dealt with *adding a new language to the ontology*, the complexity of giving multilinguality to the ontology will influence the choice of the ontology meta-model and, consequently, also the ontology model.

Summary of the main implications of WP requirements

The main conclusions extracted from this analysis can be summarized in 6 main requirements:

- 1) Ontologies have to be language independent.
- 2) Monolingual ontologies related via mappings are discarded.
- 3) Multilinguality is necessary at interface level.
- 4) Multilinguality is necessary at different ontology components. (It has to be determined exactly which ones).
- 5) Multilinguality has to be reflected in the ontology metadata.
- 6) Multilinguality has to be conferred to the ontology during the design time.

12.4 Rationale for a three layered approach and evaluation criteria

The multilingual information requirements identified in section 12.3 are to be taken into consideration when proposing Multilingual Ontology Meta-models for NeOn. Firstly, we have to make sure that the proposals we present meet those requirements, and, secondly, we will have to evaluate the advantages and limitations of the proposed meta-models, in order to come up with the solution that better meets NeOn's purposes. In a Knowledge Based Application, multilinguality has to be defined at the layers or levels in which it can appear, namely:

- 1) Interface level
- 2) Metadata level
- 3) Knowledge Representation (KR) level
- 4) Data level⁶⁴

12.4.1 Evaluation criteria for interface level

The evaluation criteria used to analyse advantages and disadvantages of multilingual interfaces are related to:

- a. **Retrieval time of multilingual queries**
- b. **Changes in the visualizing code when adding a new language**

12.4.2 Evaluation criteria for metadata level

Advantages and disadvantages of possible modifications at the Metadata level to express multilinguality have to do with:

⁶⁴ As mentioned in Note 1, the data level has not been yet included at this stage.

- a. **Quantity of linguistic information to include at the metadata level**
- b. **Capability of the system to work with relations associated to linguistic information**

12.4.3 Evaluation criteria for KR level

The criteria used to evaluate advantages and limitations at the KR level have to consider many factors, which are listed below.

- a. **Number of meta-models of the KR System**
- b. **Number of models of the KR System**
- c. **Number of Reasoners (R):** it depends on the number of models of the KR. We have identified 3 types of reasoners.
 - Ontology Reasoner (OR)
 - Mappings Reasoner (MR)
 - Linguistic Resource Reasoner (LRR)
- d. **Complexity of the query:** the level of complexity of the query is inferred from the number of models and the number of model components in the KR that have to be consulted to obtain a result. Thus, we have identified 4 different levels of complexity for the purposes of this survey, which range from 1 (lowest level of complexity) to 4 (highest level of complexity):

Level of complexity	1 model	2 models
1 component	1	3
2 components	2	4

- e. **Complexity by adding a new language:** the grade of complexity we encounter when adding a new language to the KR depends on the target element of the modification, and ranges from 1 (less complexity) to 3 (most complexity):

Target of modification	Level of complexity
metamodel	3
n models	2
1 model	1

- f. **Number of managers:** the number of managers corresponds directly with the number of models of the KRS, i.e., Ontology model, Linguistic Resource model and Mappings model.

- g. Complexity levels of consistency:** maintenance of consistency in a KRS depends on the number of managers the system needs. The number of managers is in turn dependent on the number of models that make up the KR System. The more managers are needed, the more difficult it will be to maintain consistency (being 3 the highest level of complexity).

Complexity	Level of complexity
c (constant)	1
n	2
n²	3

- h. Real availability:** this criterion indicates if all components (reasoners, managers) are currently available or not.

13. Representation of multilinguality at the Interface level

The interface can support multilinguality at two different levels:

- message level
- content level

According to the identified requirements, the message level of the interface has to be multilingual, as well as the content level, i.e., the information requested by the user has to be displayed in the desired language.

13.1 Multilingual interface at message level

Multilinguality can be in turn represented in two ways, by presenting the messages of the page in multiple languages simultaneously (Figure 37), or by giving the option of alternatively visualizing the interface in the different languages, one at a time (Figure 38).

Figure 37: Simultaneous multilingual interface messages

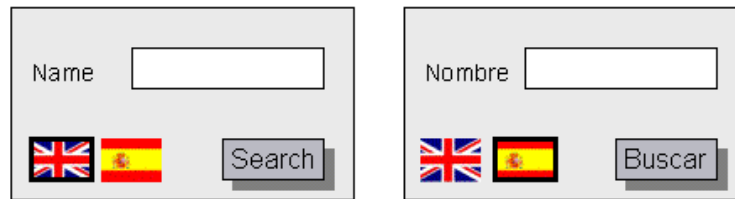


Figure 38: Alternately monolingual interface messages in a multilingual system

13.2 Multilingual interface at content level

In order to allow the user to access content written in different languages, two options can be applied:

- a) the KB is **multilingual** and, therefore, the information retrieval takes place in the selected language;
- b) the KB is **monolingual**, but by accessing a lexical resource which enables a translation process, the information is presented in the selected language.

However, as stated in the requirements, multilinguality has to be conferred to the ontology in the design time, so the b) option has to be discarded.

13.3 Advantages and disadvantages of a multilingual query

One advantage of a multilingual KB is that the retrieval time of a query is equivalent to the reply time of the KB, since the localization process has been carried out during the **design time** of the KB. However, the localization process is more time consuming, since translation problems, for example, disambiguation problems, have to be solved while developing the multilingual application.

13.4 Advantages and disadvantages of adding a new language to the interface

Depending on the kind of multilingual interface chosen for our application, certain requirements will have to be addressed when adding a new language to the interface.

- a) If multilingual messages are displayed simultaneously (Figure 37), the whole existing visualizing code will have to undergo modifications when adding a new language.
- b) If multilingual messages are displayed alternately, no modification of the visualizing code is needed, but the number of interfaces will increase, as well as the number of elements that represent the languages in which applications are available, i.e. flags in Figure 38.

Consequently, we regard the b) option as the more appropriate one for representing multilinguality in NeOn.

14. Multilinguality in a Knowledge Representation System (KRS)

Multilinguality in a Knowledge Representation (KR) has a three-fold perspective: **Information**, **Modeling** and **Realization**. The **Information** aspect in a KR refers to the metadata that gives some kind of global information about the ontology, for example, information about the authorship or creation date of the ontology. As established in D1.1.1, the standard for ontology metadata used in NeOn is the Ontology Metadata Vocabulary, also known as OMV (Hartmann *et al.* 2006). **Modeling** has to do with the representation of the components of the KR, i.e. the representation of the ontology meta-model. **Realization** is the real expression of the meta-model in the KR, i.e. the ontology model. The latter two aspects are included in the so-called Knowledge Representation level.

14.1 OMV level: Modification/Extension of the Core Model

In a multilingual KR, information about multilinguality should be part of the metadata of such a representation. Therefore, next to conceptual metadata such as authorship of the model, creation date, or engineering tool used in its development, we should find information about multilinguality. First, at the metadata level, information about the natural language(s) in which the representation system is available (in this particular case, the ontology) should be enough. However, as will be shown further in this document, it may also be necessary to express which components of the (ontology) representation support multilinguality or even which kind of linguistic data they do support.

In order to store this kind of information, we have identified two possible methods exemplified in Figure 39 and Figure 40, by adding an extension to the OMV Core or by modifying it:

- 1) We could extend the OMV by creating, for example, an **OntologyComponent** class that allows us to talk about the different elements of an ontology (Classes and Properties in an ontology following the DL paradigm), and an **OntologyNaturalLanguage** class related to the first one by means of an "is expressed in" relation, that says that a certain component of the ontology is expressed in a certain language. In this way the ontology metadata can represent which components of the ontology are expressed in a certain natural language. Let us say, for example, that in an ontology, Classes are expressed in English and Spanish, and Properties or Axioms just in English.

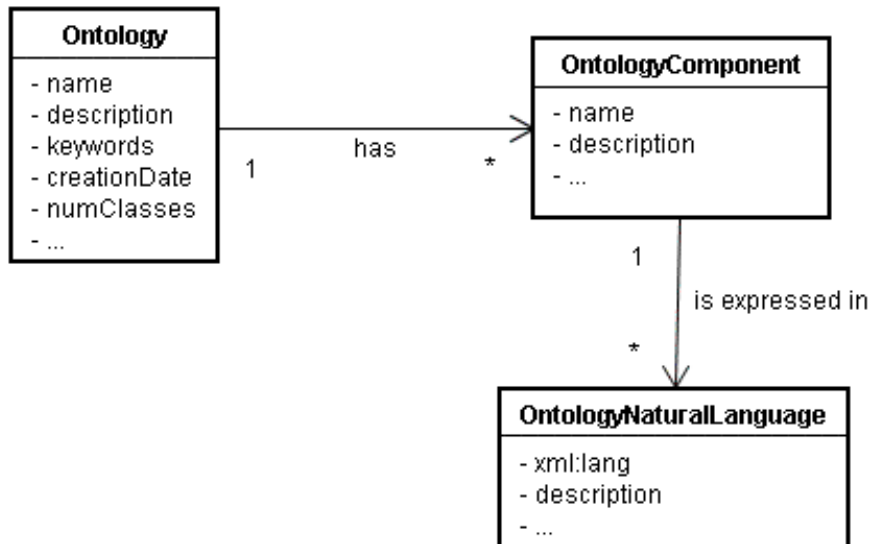


Figure 39: Example of a possible extension to the OMV

- 2) We could also modify the OMV by adding an attribute to the **Ontology** class, which could be a tuple with multiple values, composed of the ontology component and the natural language in which that component is expressed, i.e, we would merely say at this metadata level that certain components of the ontology are expressed in one or more natural languages.

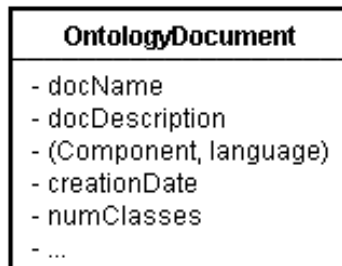


Figure 40: Tuple with multiple values about linguistic information in OMV

Both methods would solve the problem of expressing information about multilinguality, and in both cases, this representation would be independent from the approach followed for the KR.

14.1.1 Advantages and disadvantages of both representation systems

Although at a first glance both methods would solve the problem of expressing that some ontology elements have linguistic information associated in different languages, the second method is more limited than the first one.

The **OMV modification** represented by Figure 40 would just report about a certain ontology element being expressed in various natural languages, let us say, Properties are expressed in English and in German.

The **OMV extension** represented by Figure 39 allows the addition of as much information as necessary in order to report about the multilingual aspects of the ontology. By adopting this solution we would be able to report not only about the natural languages in which a certain ontology element is expressed, but also about the sort of linguistic data associated to it. For example, it could be said that lexicalizations and definitions in English and German are associated to Properties.

The point here is to weigh up how much linguistic information is relevant for the application. The agreed solution will also depend on the Multilingual Ontology Meta-model defined for NeOn.(Cf. section 16.2)

14.2 KR level

The next aspect to be considered in a KRS is the modelling of the meta-model. Modelling multilinguality in ontologies can result in different representation forms, depending on the modifications undergone by the meta-model. Taking into account the requirements analysed in **section 12.3**, we have identified two modelling possibilities to represent multilingual meta-models, namely:

1st Proposal - Modified Ontology Meta-model: multilinguality is embedded in the NeOn meta-model

2nd Proposal - Ontology Meta-model linked to a Linguistic Information Repository (LIR) Model⁶⁵: multilinguality is not embedded in NeOn meta-model

The first proposal is based on a modification of the ontology meta-model by adding multilinguality to it. The second proposal involves the creation of an independent Linguistic Information Repository (LIR) model that is then related to the ontology meta-model. Both proposals result in the creation of a **Multilingual Ontology Meta-model (MOM** from now on).

14.2.1 Modified Ontology Meta-model

The first identified meta-model can support multilinguality by modifying the ontology meta-model. In this approach, depending on which ontology elements support multilinguality, there will be different modification levels of the ontology meta-model.

Figure 41 shows an extract from the OWL DL meta-model adopted for NeOn, in which **Classes** and **Properties** are represented (D1.1.1: 25). We will use the upper part as basic representation of the NeOn meta-model for exemplifying the meta-model representations identified in this section.

⁶⁵ Note that what we call Linguistic Resources (LRs) is referred to as Knowledge Organization Systems (KOS), in D1.1.1, page 15, and they are considered *important sources for ontology construction*. Even the creation of *meta-models to map a KOS meta-model to OWL meta-model* is pondered.

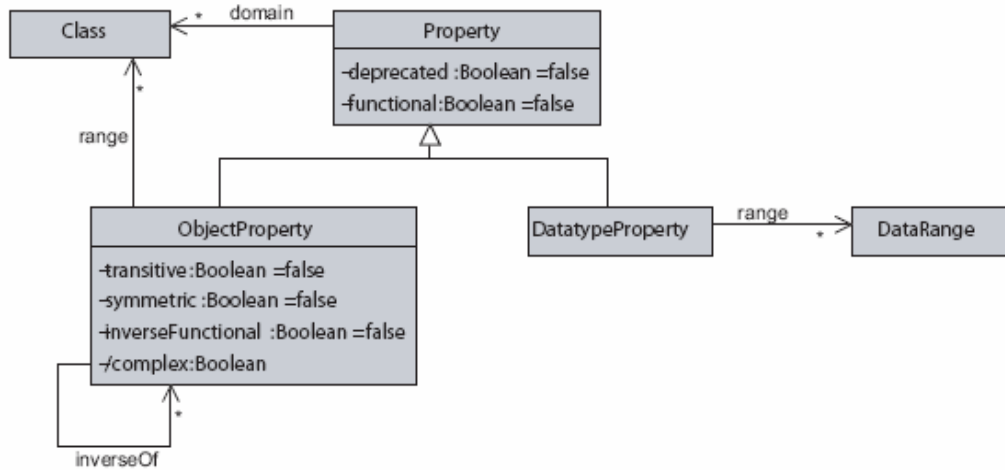
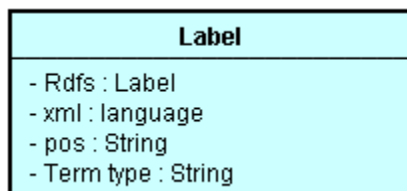


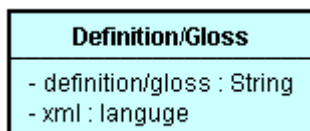
Figure 41: Classes and Properties of the OWL DL Meta-model for NeOn (D1.1.1: 25)

Let us assume that the linguistic information we want to add to our ontologies in NeOn consists of:

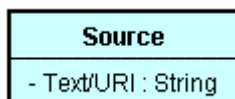
- 1) Lexicalizations of ontology elements in the different languages of our resource, what we have called `Label`.



- 2) An explanation or definition of the ontology element in natural language called `Definition/Gloss`

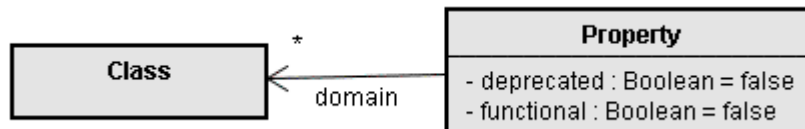


- 3) A class named `Source`, in which the source from where the linguistic data above mentioned has been extracted is reported.



Now let us assume that our ontology meta-model consist of 2 ontology elements:

- 1) Class
- 2) Property



The modification of the ontology meta-model consists of the addition of three new classes to the ontology meta-model, namely *Label*, *Definition/Gloss* and *Source*, which will be linked to the ontology classes *Class* and *Property*, as can be seen in Figure 42. According to the requirements analysed in **section 12.3**, the *user is requested to specify which elements of the ontology should be multilingual: user can select **classes, instances, properties** or the **entire ontology*** (Use case 4.6.15). Therefore, the proposed MOM in which multilinguality is embedded in the NeOn meta-model would meet the requirements.

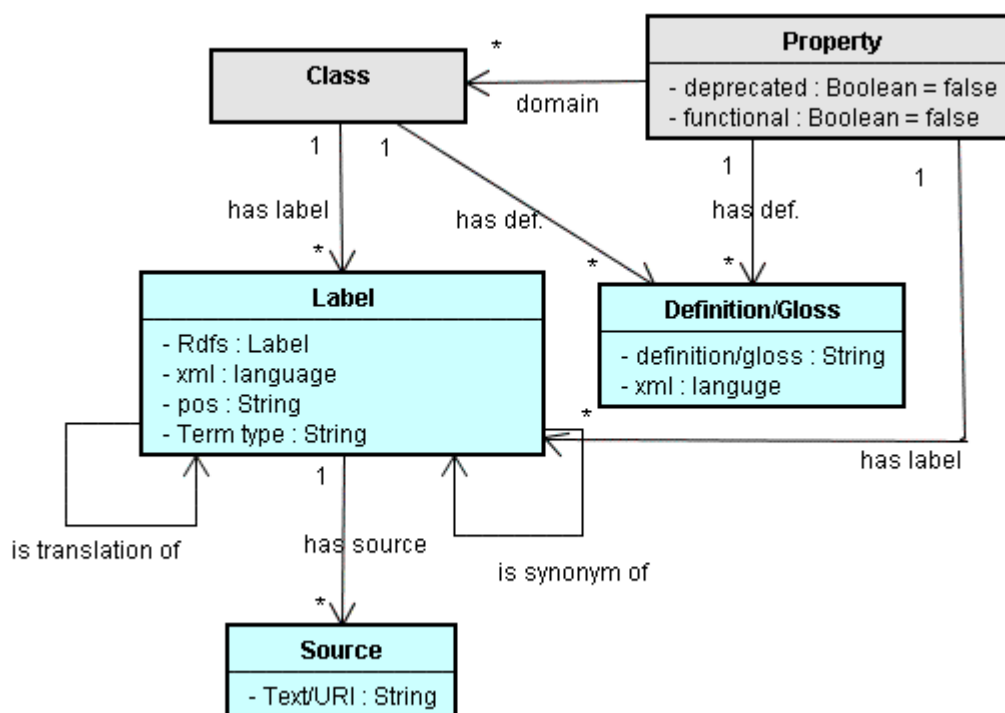


Figure 42: MOM represented by Label, Definition/Gloss and Source classes linked to Class and Property

14.2.2 Advantages and disadvantages of a Modified Ontology Meta-model

According to the evaluation criteria identified in **section 12.3**, the number of meta-models (a) would be only 1, and therefore, the number of reasoners (c) and managers (f) is also 1. This means that consistency (g) is not difficult to maintain. The level of complexity of the query (d) is 2, because in order to solve a query, the system needs to consult one model and two or three components. When adding a new language (e) no complexity would be involved, since the meta-model does not need to be modified. However, the problem of having the ontology in so many languages could result in difficulties to manage it, because the amount of components would considerably increase.

A summary of the advantages and disadvantages has also been included below:

- **Advantages:**

- The ontology representation would be independent from the language information (the so-called language layer)
 - Complexity in adding a new language would be low, because no meta-model modification is necessary
 - Complexity in maintaining consistency would also be low, because there is just one ontology model to be managed
 - The tools or systems required already exist
- **Disadvantages:**
- Complexity in the process of the query is high, since the ontology has many components to manage
 - If more linguistic information is to be added, the amount of ontology elements can be very high

14.2.3 Ontology Meta-model linked to a Linguistic Information Respository (LIR) Model

The second approach for creating a MOM consists of a **Language-independent Ontology Meta-model** and a Linguistic Information Repository (LIR) model, linked to each other. On the one hand, there is the ontology meta-model in which multilinguality is not considered, and, on the other hand, a **LIR** model is modelling some specific linguistic information, as shown in Figure 43. The so called LIR model consists of three elements: Label, Definition and Source.

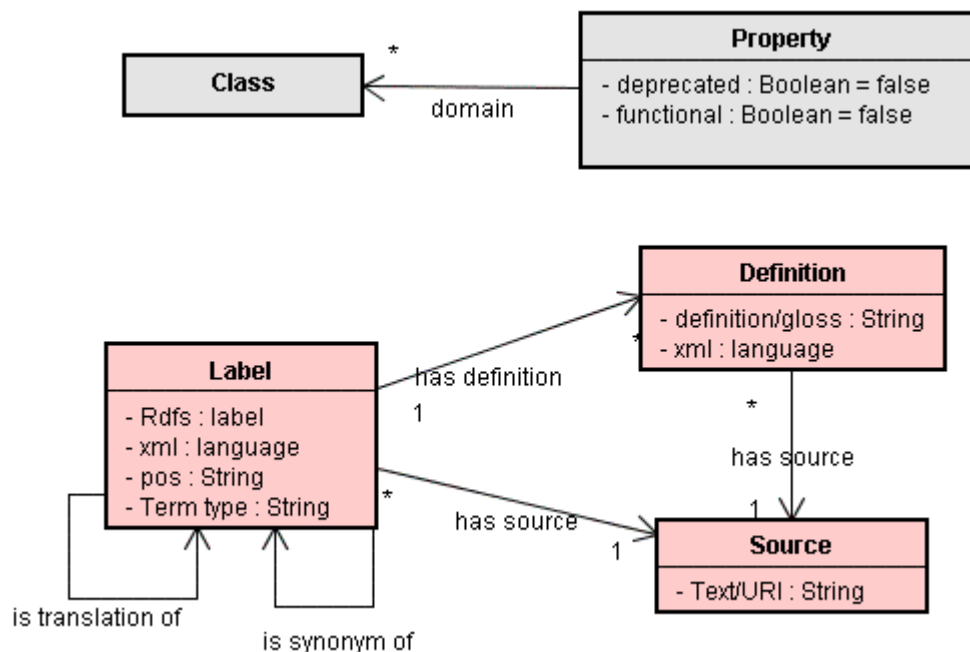


Figure 43: Two Models: the Ontology Meta-model and the LIR Model

As illustrated in Figure 44, all ontology elements (**Classes** and **Properties** in our example) would be linked to the LIR model through the `Label` class, thus conferring multilinguality to the system.

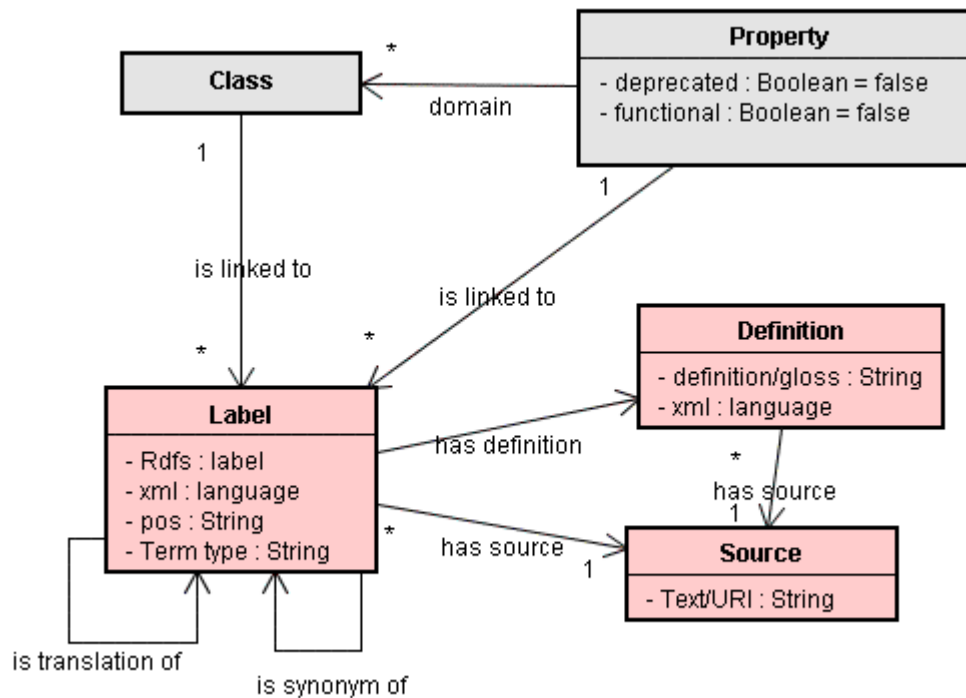


Figure 44: Example of a MOM represented by a LIR Model linked to the Ontology Meta-model

14.2.4 Advantages and disadvantages of an Ontology Meta-model linked to a LIR Model

A comparison of this meta-model against the evaluation criteria identified in **section 12.4** gives the following results: this KR is formed by 2 meta-models (a), which imply the existence of 2 reasoners (c) and 2 managers (f). The level of complexity of the query (d) is thus 3.

The grade of complexity created by adding a new language (e) to this system is 1 (the lowest level of complexity), since it does not imply a modification of the ontology meta-model. In the same way, the criterion of real availability (g) will be satisfied by the representation system chosen for the LIR. However, it is highly probable, that the LIR acquires the form of an ontology, so it would make use of the tools and systems that manage the ontology.

We have identified the following advantages and disadvantages:

Advantages:

- The ontology representation is independent from languages (the so called language layer)
- Complexity by adding a new language is low, because no meta-model modification is necessary
- If the LIR is modelled as an ontology, the use of tools, systems and access mechanisms already defined for NeOn ontologies can be reused.

- The LIR can contain as much linguistic information as the user wishes without interfering or creating noise in the ontology
- The LIR model can be compliant with already established linguistic representation standards allowing the re-use and sharing of data

Disadvantages:

- The level of complexity in the query would be mid-scale, because of the existence of 2 models (which implies the need of 2 reasoners and 2 managers)
- Complexity in maintaining consistency would be mid-scale because two models have to be managed

15. Ontology Models: realization and instantiation

The realization of multilinguality in a KRS is strictly related to its modelling. These two levels are intrinsically interrelated -the one cannot exist without the other- since the realization is nothing else but an instance of the modelling, and the model in turn is an instance of the meta-model.

There are different realizations depending on the meta-models we have identified in the previous section, and which have been grouped as follows:

1st Proposal – Realization of the Modified Ontology Meta-model: Multilingual Ontology Meta-model

2nd Proposal – Realization of the Ontology Meta-model linked to a Linguistic Information Repository Model⁶⁶: Ontology Meta-model linked to a LIR Model

15.1 1st Proposal: Realization of the Modified Ontology Meta-model

In this section we include those realizations which correspond to the approach followed by our 1st Proposal: Modified Ontology Meta-model. This approach allows the inclusion of the necessary multilingual information in the ontology.

As identified in the previous section, there are different possibilities in the modification of the ontology meta-model in order to include linguistic data that allows the representation of multilinguality. In Figure 45 we can see an example of the Ontology Model corresponding to the Ontology Meta-model presented in Figure 42.

In this case, the `Class` is associated to as many `Label`, `Definition` and `Source` instances as languages are considered in the ontology. In the example below, instances in English.

⁶⁶ Note that what we call Linguistic Resources (LRs) is referred to as Knowledge Organization Systems (KOS), in D1.1.1, page 15, and they are considered *important sources for ontology construction*. Even the creation of *meta-models to map a KOS meta-model to OWL meta-model* is pondered.

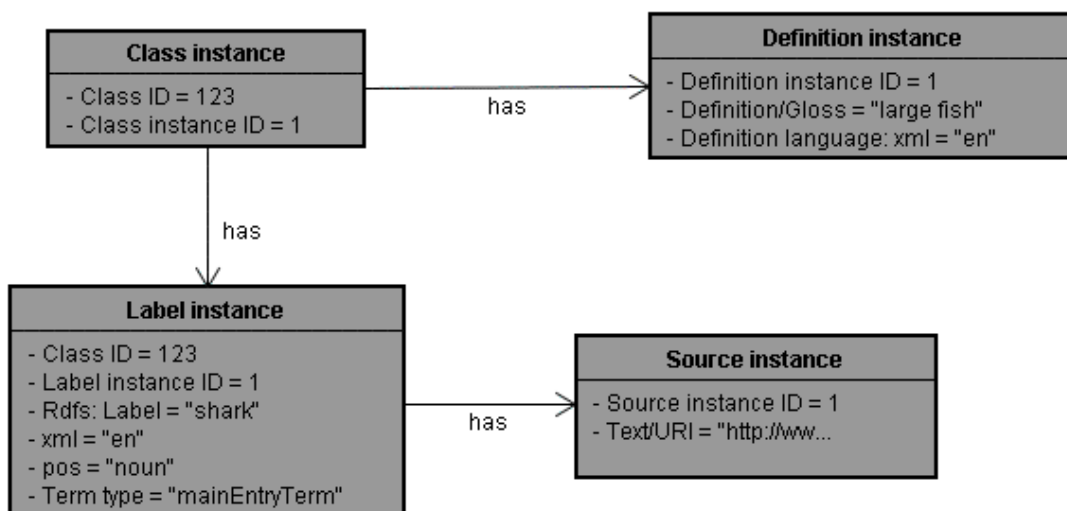


Figure 45: Example of an Ontology Model based on a Modified Ontology Model with multilingual Instances associated to Classes

15.2 2nd Proposal: Realization of the Ontology Meta-model linked to a LIR Model

On the assumption that the linguistic information has its own entity and turns into an independent Model apart from the Ontology Meta-model, we can speak about multilingual models that include an **Ontology Meta-model**, a **Model of the Linguistic Information**, and **links between both models**. The linguistic resource could be a relational database or an ontology, for example. As already mentioned for the case of the NeOn ontologies, should we represent the linguistic information by means of an ontology, then those tools already available for NeOn could be reused.

The realization of the exemplified Figure 44 of the Ontology Meta-model linked to the LIR Model could look like the figure below with instances in English and Spanish for ontology Classes.

The attributes for each class are for illustration purposes only. For an exhaustive description of the linguistic and terminological coverage of the model please see chapter 16.

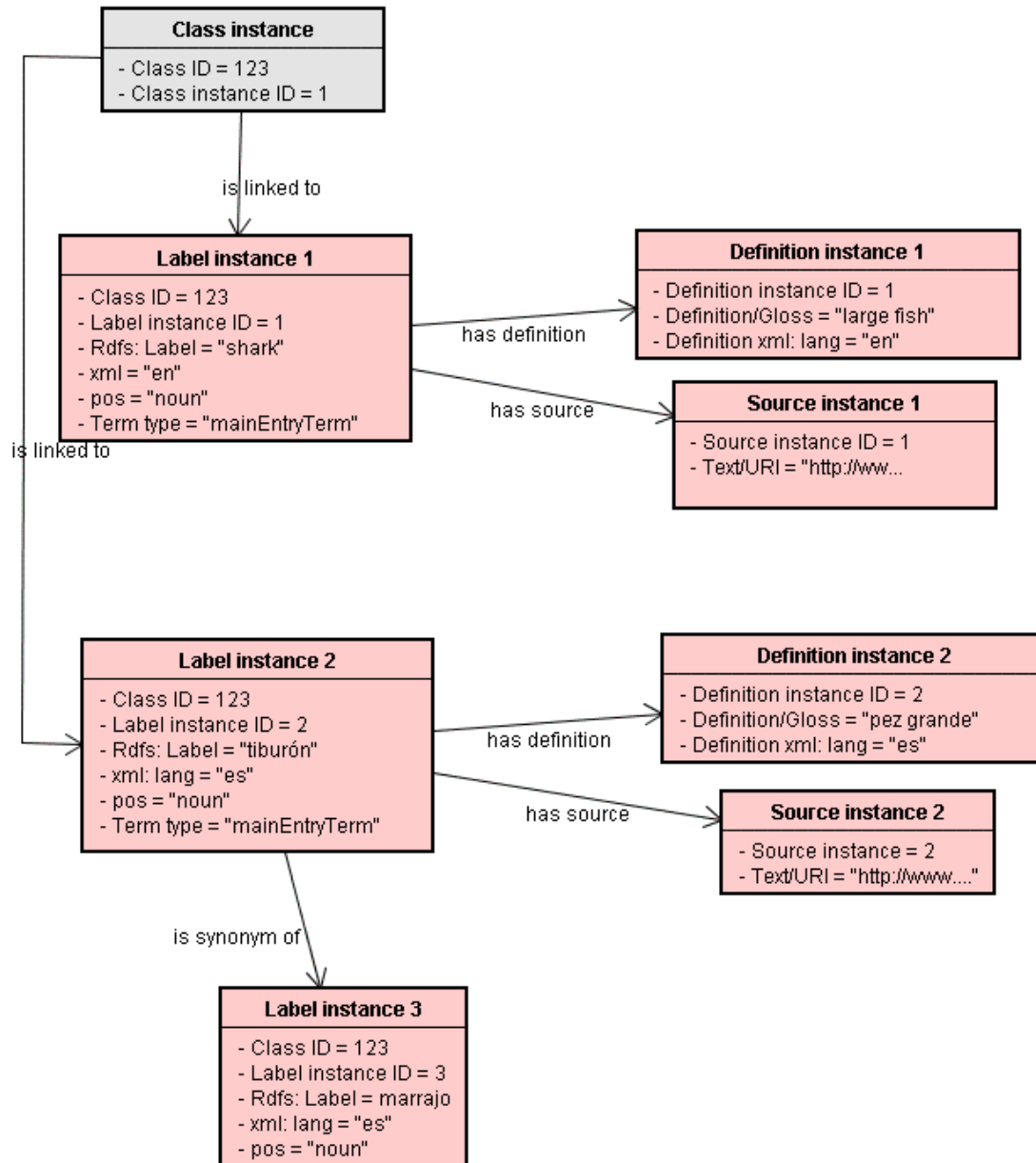


Figure 46: Example of an Ontology Model linked to Multilingual LIR Instances

A real example of an ontology that follows the representation system in Figure 44 is GENOMA-KB⁶⁷ (Cabré 2004), which has been broadly described in **section 10.4**. The GENOMA Knowledge Base System consists of an ontology of *Concepts* or *Classes* -not associated to any language- with links to a Terminology Base. This Terminology Base contains the whole linguistic information –English, Spanish, and Catalan entries- which is associated to the *Classes* of the ontology, and, therefore provides multilinguality to the system. No terminology entry can be added to the ontology, unless the *Class* has been previously introduced in the ontology.

⁶⁷URL:

<http://genoma.iula.upf.edu:8080/genoma/corpSearch.do;jsessionid=C5F6DA7C2954A5084D48F35666F8B0DE?operation=init>

15.3 Hybrid systems

After having analysed possible ways of representing multilinguality at the three different levels identified in a Knowledge Based Application, i.e. **Interface**, **KRS** and **Meta Data**, we would like to highlight the feasible combinations of different multilingual systems in the same Application, in what we have called *Hybrid systems*.

As already mentioned, each component in an ontology model is able to support multilinguality. However, it is also possible to provide multilinguality to several components following different systems.

Let us assume that we have a multilingual application in which we have provided multilinguality to `Classes` by modifying the Ontology Meta-model, but regarding `Properties`, we have decided to do it by associating the Ontology Meta-model to a LIR Model. In this case, we are just combining two different approaches of multilingual meta-models, respectively analysed in sections **14.2.1** and **14.2.3**, in order to obtain a Multilingual Ontology Meta-model in three ontology components. We could even confer multilinguality to `Classes` in **the design time**, and to `Properties` in **the run time**, and still have a multilingual system.

Table 17: Criteria for identifying advantages and limitations of MOM

Multilingual Ontology Meta-model (MOM)	Multilingual Ontology Realization (MOR)	Number of meta-models of KR (a)	Number of models: ontology model (O), and LR model(b)	Number of Reasoners (c)	Complexity in the query (d)	Complexity by adding a new language (e)	Complexity in maintaining consistency (g)	Real availability (h)
MOM represented by a modified Ontology Meta-model – Figure 42	Figure 45	1	1 (O)	1 OR	2	1	1	YES
MOM based on an Ontology meta-model linked to a LIR model – Figure 44	Figure 46	2	1(O) (LIR) + 1	1 OR 1 LRR	3	1	2	YES

16. The Multilingual Ontology Meta-model proposed for NeOn

16.1 NeOn Ontology Meta-model linked to the LIR Model

16.1.1 The choice of Model

After careful consideration of the requirements for the NeOn linguistic model, and the various options for organizing this information (see previous sections), the authors recommend to adopt the separation of ontological and linguistic information, i.e. the Linguistic Information Repository described in the second proposal (see section 14.2.3). According to this model, conceptual and linguistic information is captured in different modules of the NeOn framework:

1. The ontology meta-model as defined in D1.1.1
2. A linguistic/terminological meta-model, called the Lexical Information Repository (LIR) , which captures all the relevant linguistic/terminological information associated with concepts such as lexicalizations, lexicalization types and multilinguality.

This modular approach to the overall meta-model architecture ensures separation of information that is considered orthogonal in nature.

On the one hand, ontologies are conceptual constructs without linguistic content. From a formal ontological point of view, concepts are abstract notions whose labels are arbitrary. On the other hand, the orthography and senses of the lexicalizations that function as labels for these concepts are only considered to be evocative or indicative of the ontological meaning of the concepts. There is an implicit mapping assumption between lexical and conceptual knowledge, which underlies "ontology lexicalization", namely that (intensional) senses from a lexical model are mapped to (extensional) interpretations on ontology elements (classes, properties, individuals, restrictions). The lexical semantic content of the lexicalizations, originating from linguistic/terminological resources such as term banks, thesauri and dictionaries, is considered to be lightweight, and in need of formalization in order to become a fully-fledged ontology.

In order to capture and represent the interplay between conceptual and lexical meaning, we need to define a model which links both types of meaning by means of an ontological module on the one hand, and a linguistic/terminological module on the other.

The linguistic/terminological meta-model in Figure 47 below has been designed from the perspective of the ontology engineer. It takes relevant linguistic and terminological knowledge from resources into account, such as term banks, thesauri and dictionaries, in order to create a linguistically/terminologically informed link between intra- and extra-ontological information.

It is a structured, non-exhaustive set of linguistic and terminological data categories, built up on the basis of existing standards. This ensures interoperability with these standards, and a maximum level of acceptance within the user communities, active in the combined fields of linguistics, terminology and ontology engineering.

It is extensible in the sense that it will be able to accommodate any additional data categories deemed useful for an ontology engineer editing lexicalizations and browsing available linguistic information such as alternative lexicalizations and translations. For instance, the class `UsageContext` (see Figure 47 below) can be extended with new subclasses from the TBX data category proposal⁶⁸, such as definitional and associative context. Also, further morphological and

⁶⁸ <http://www.lisa.org/standards/tbx/>

syntactic decomposition such as headword identification and stemming can be included (Buitelaar *et al.* 2006). Moreover, foreseeable future developments, such as a typology of definitional structure, can be added without the stamp of official standardization, while still building on standard information structures.

The association between the OWL meta-model and the LIR is established by the `hasLexicalEntry` relation between `OntologyElement` and `LexicalEntry`. The latter manages the access to the linguistic and terminological knowledge.

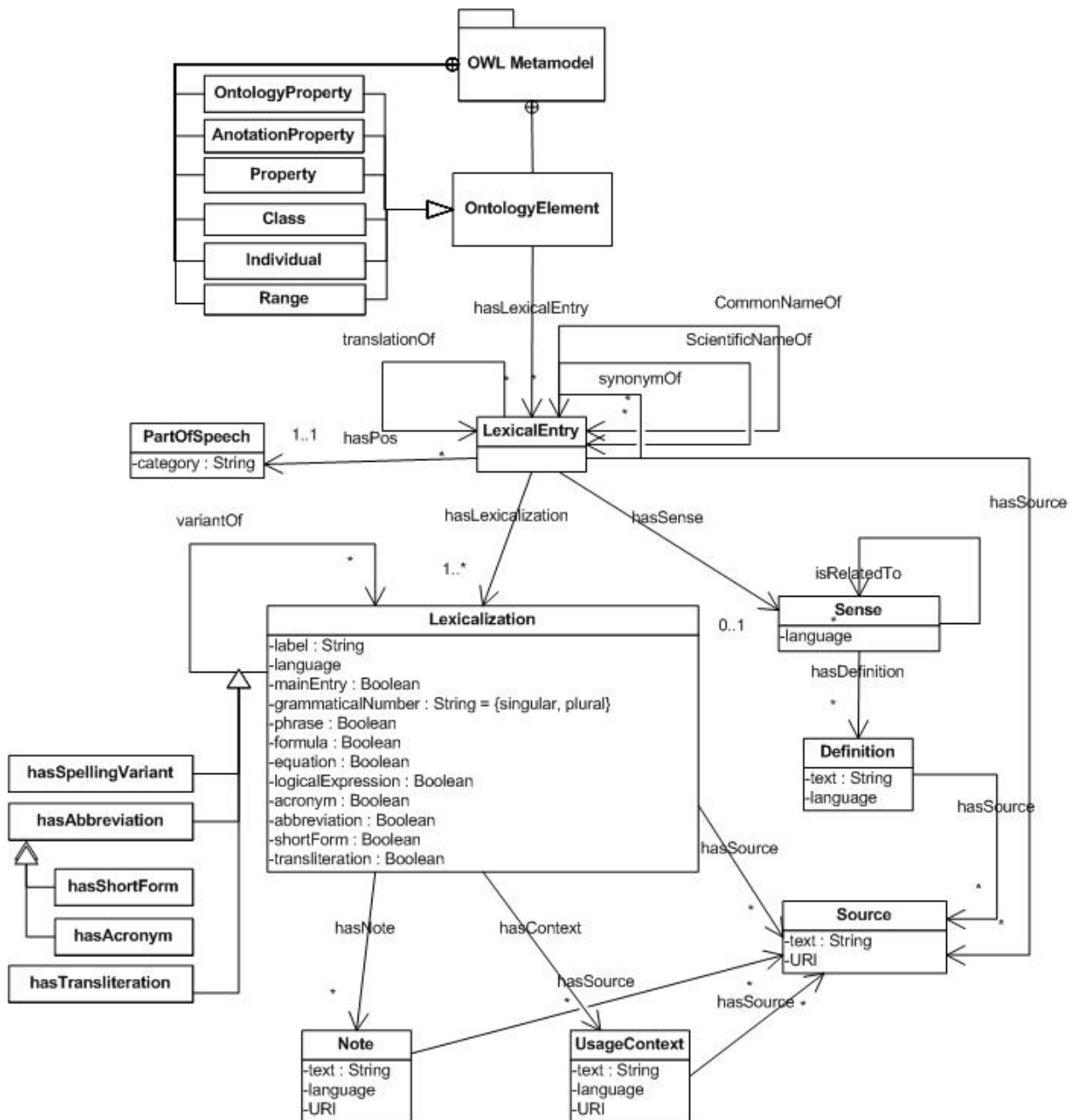


Figure 47: The LIR model

16.1.2 Description of the classes:

1. LexicalEntry: a lexeme, which is a unit of form and meaning.

A lexeme is an ordered collection of related word forms, having the same lexical meaning⁶⁹ (Saloni *et al.* 1990).

Please note that the meaning shared by the word forms is lexical, not grammatical. Meaning differences between e.g. singulars/plurals are not covered by lexical meaning.

The `LexicalEntry` class manages the link between sense and lexicalization. It is an abstract class, of which each instance is a combination of a set of `Lexicalizations` and zero or one sense.

2. Sense: a language-specific unit of intensional lexical semantic description.

The addition of the attribute `xml:lang` to `Sense` allows us to model language specific meaning. `Lexicalizations` in multiple languages can be linked in the following ways:

- they are offered to the user as translation pairs without any indication of sense;
- they are associated with the same sense or terminological entry (as in TMF and TBX),
- they each have their own language specific sense with sense relations such as cross-language synonymy or e.g. near-synonymy (either pair-wise between languages, or through an interlingua).

In order to be able to capture language-specific aspects of meaning, all `Lexemes` are language specific, and their translational or conceptual equivalence is expressed by the relations `hasTranslation` and `hasSynonym`.

3. PartOfSpeech: The grammatical class of the `LexicalEntry`.

Traditionally, members of the set of word forms incorporated into a particular lexeme are selected on the basis of part of speech, inflectional behaviour and meaning.

The fact that lexemes are pre-filtered by syntactic class means that adding `PartOfSpeech` to `LexicalEntry` avoids repetition of `PartOfSpeech` for all `Lexicalization` instances. Synonymy relations across part of speech boundaries will need to be implemented at the `LexicalEntry` level.

4. Lexicalization: a word form

This class corresponds with the LMF class `Form Representation`. The choice of this data category means that the lexicalizations of concepts are deemed word forms rather than lemmas or citation forms, and are therefore allowed to be inflected forms, such as plurals.

The notion of `Lemma` as the canonical form (citation form) representing the set of related word forms such as inflections, is equivalent to `Lexicalization` with attribute `mainEntry` (see below) set to `true`.

The class `Lexicalization` has the following attributes:

- `Rdfs:Label`
- `Xml:lang`: language code from ISO639-2⁷⁰

⁶⁹ See also Wikipedia: <http://en.wikipedia.org/wiki/Lexeme>

- *GrammaticalNumber* captures the lexicalization's morpho-syntactic features such as plurality and singularity.

Further, it contains a set of descriptions for term types taken from TMF⁷¹ and TBX-Lite⁷², split up into:

A. a set of Boolean attributes describing a number of term types:

- *mainEntry* (The concept designation that has been chosen to head a terminological record.) (ISO12620: section 02.01.01)
- *Formula* (Figures, symbols or the like used to express a concept briefly, such as a mathematical or chemical formula) (ISO12620: section 02.01.14)
- *Equation* (An expression used to represent a concept based on a statement that two mathematical expressions are, for instance, equal as identified by the equal sign (=), or assigned to one another by a similar sign) (ISO12620: section 02.01.15)
- *Symbol* (A designation of a concept by letters, numerals, pictograms or any combination thereof) (ISO12620: section 02.01.13)
- *LogicalExpression* (An expression used to represent a concept based on mathematical or logical relations, such as statements of inequality, set relationships, boolean operations, and the like.) (ISO12620: section 02.01.16)
- *Phrase*: A phraseological unit containing any group of two or more words that are frequently expressed together and that comprise more than one concept. The individual words in a phrase usually function in more than one grammatical category (part of speech) within the syntax of a sentence.e.g. "work offline") (ISO12620: section 02.01.18)
- *ScientificName*: A term that is part of an international scientific nomenclature as adopted by an appropriate scientific body. (ISO12620: section 02.01.04)
- *Acronym*: An abbreviated form of a term made up of letters from the full form of a multiword term strung together into a sequence pronounced only syllabically. (ISO12620: section 02.01.08.04)
- *ShortForm*: An abbreviated form that includes fewer words than the full form. e.g. "Intergovernmental Group of Twenty-four on International Monetary Affairs" vs. "Group of Twenty-four". . (ISO12620: section 02.01.08.02)
- *Abbreviation*: A term resulting from the omission of any part of the full term while designating the same concept, e.g. adjective vs. adj. (ISO12620: section 02.01.08)
- *Transliteration*: A form of a term resulting from an operation whereby the characters of an alphabetic writing system are represented by characters from another alphabetic writing system. (ISO12620: section 02.01.10)

B. a number of relations between `Lexicalization` classes:

- `hasSpellingVariant`
- `hasAcronym`
- `hasShortForm`
- `hasAbbreviation`
- `hasTransliteration`

⁷⁰ http://www.loc.gov/standards/iso639-2/php/English_list.php

⁷¹ <http://www.ttt.org/oscar/xt/webtutorial/datcats02.htm>

⁷² <http://www.lisa.org/standards/tbx/>

`hasAcronym` and `hasShortForm` are subtypes of `hasAbbreviation`. Although both have been officially disallowed, and the use of the more general attribute *Abbreviation* is prescribed, FAO requires these data categories.

`hasScientificName` and `hasCommonName` have been defined as relations between `LexicalEntries`. This gives us a more economical representation of this information, because it reduces the reduplication of this information at the `Lexicalization` level. If we maintain the `hasScientificName` relation as a relation between `Lexicalizations`, we need to encode this relation between each common name `Lexicalization` within each `LexicalEntry` and each `ScientificName Lexicalization`, not only within a language, but also across languages, since the `ScientificName` is the same for each language specific `CommonName`.

In many cases, the directionality of these relations enables the derivation of term types as Boolean attributes for `Lexicalization` classes. For instance:

$X \text{ hasScientificName } Y \rightarrow X: \text{ScientificName}: 0; Y: \text{ScientificName}: 1$

$X \text{ has Abbreviation } Y \rightarrow X: \text{fullForm}; Y: \text{Abbreviation}$

$X \text{ hasSpellingVariant } \rightarrow X: \text{mainEntry}; Y: \text{Variant}$

Representing these term types as relations rather than as Boolean attributes ensures the proper link between unique source and target lexicalizations.

The reason for using both a set of Boolean attributes and a set of relations is that relations cannot always be deduced from a set of attributes. For instance, if two lexicalizations are associated with a concept, and one of them is an abbreviation, then it is impossible on the basis of Boolean attributes to determine if the full form lexicalization is related to the abbreviation.

Conversely, attributes cannot always be deduced from relations, in cases where there is only one word form as lexicalization.

5. Definition: A statement that describes a concept and permits its differentiation from other concepts within a system of concepts. (ISO12620: section 05.01)

The Definition class has the following attributes:

1. Definition/Gloss: string.
2. Definition language: `xml:lang`

6. Source: the provenance of the linguistic/terminological information. This can be expressed by the following data categories:

1. a name space identifier (ISO12620: section 10.21),
2. a bibliographic reference: A complete citation of the bibliographic information pertaining to a document or other resource. (ISO12620: section 10.19)
3. a source identifier: The code assigned to a document in a terminological collection and used as both the identifier for a bibliographic entry and as a pointer in individual term entries to reference the bibliographic entry identified with this code. (ISO12620: section 10.20)

The name space constitutes a unique index into the source resource, and is therefore the preferred attribute. If this is not available, the external link can be expressed in the `Text` attribute, e.g. by means of the URL of the resource, a textual description of the resource, or maybe a unique key

into the resource specific information structure (for instance, in the case of a dictionary, the composite key lemma, pos and sense number).

7. UsageContext: A text or part of a text in which a term occurs. (ISO12620: section 05.03)

A more fine-grained typology of context is expected within ISO, with subcategories such as `definingContext`, `explanatoryContext`, `associativeContext` and `linguisticContext`.

8. Note: Supplemental information pertaining to any other element in the data collection, regardless whether it is a term, term-related, descriptive, or administrative. (ISO12620: section 08)

This class can be linked to any element from this model, classes and properties. For instance, notes associated with the synonym and translation properties can informally describe differences in meaning between lexically synonymous labels on the one hand, and differences in meaning between translational variants on the other. For the moment, these differences are envisaged to be captured in a non-formal way through free text. It is possible that in a later stage these differences can be formalized to a greater extent.

16.1.3 Description of the relations:

1. `hasLexicalEntry` The link between ontology and LIR.

This relation has, as yet, no semantic characterization apart from “is lexicalized by”. It can be further parameterized in order to describe the nature of the mapping between lexical and conceptual knowledge. For instance, an element from a lightweight ontology can be linked to an LIR `LexicalEntry` with conceptual equivalence.

The ontology engineer decides if a lexical entry applies to a concept to a sufficient level of satisfaction. If we consider, for instance, the use of semantic and conceptual features in the description of concept and sense, it is possible to create a further sub-classification of this correspondence relation along the lines of set relations such as subset, overlap, and even disjointness (Holi and Hyvönen 2004).

2. `SynonymOf`: lexical semantic equivalence relation between `LexicalEntries`.

The decision whether two `LexicalEntries` in different languages are synonyms depends on the characterization of the synonymy relation. Since labels are elements from natural language, the logical notion of synonymy (the preservation of truth conditions in all contexts) is hardly ever applicable.

Because of this fact, Miller and Fellbaum (1990) suggest to use a weaker notion of synonymy, namely 'semantic similarity', which is defined as:

“two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value” (Miller *et al.* 1990).

So synonymy within one context is used in WordNet as the criterion for putting two lexemes together in one synset.

3. `TranslationOf`: lexical semantic equivalence relation between `LexicalEntries` from different languages.

4. `hasSpellingVariant`: a relation between `Lexicalizations` describing variance in orthographic representation.

5. Both **hasAbbreviation** and **hasTransliteration** are relations between `Lexicalizations`, and related to the attributes `Transliteration` and `Abbreviation` described above. They are subtypes of the general **hasVariant** relation.

6. **hasAbbreviation**: also a subtype of `hasVariant`. This in turn subsumes the following relations: `hasShortForm` and `hasAcronym`, which are related to the attributes `ShortForm` and `Acronym` described above.

7. **hasNote**: relation between any `OntologyElement` and `Note`.

8. **hasSource**: associates various classes with `Source`

9. **hasDefinition**: associates `Sense` with `Definition`

10. **hasSense**: associates `LexicalEntries` with `Sense`

11. **hasPos**: associates `LexicalEntries` with `PartOfSpeech`

16.1.4 LIR properties

The units of description that have been selected for the LIR form an eclectic set of data categories. These are considered to constitute useful information for ontology engineers when e.g. editing lexicalizations and browsing available linguistic information such as alternative lexicalizations and translations.

As indicated above, the data categories are a subset of available data categories from several ISO standards, such as TMF (TBX), LMF and SKOS (see sections 2 and 12.2). This ensures a maximum level of coverage, interoperability and acceptance within the communities, brought together. Also, it avoids re-inventing the wheel, and proposing yet another model for capturing these types of knowledge.

The present set of data categories incorporated into the LIR model is fixed for the moment, but by no means rigid. It covers the FAO requirements.

The interconnectivity with existing standard models for lexical and terminological description allows any dynamic extension of the LIR for the user: any additional data category from a resource in the recognised standard representations can be accessed through extended navigation. Moreover, resources modelled in other, widely used, de facto standard representations, such as TEI and JAVADICT, can be linked up by associating their units of description with standard data categories.

In short, the flexibility and extensibility and interconnectivity make the LIR into a versatile gateway into linguistic and terminological knowledge.

Possible future integration of other standards, such as MLIF (see section 12.2) can be easily envisaged within this architecture.

16.2 OMV extension for capturing multilinguality: LexOMV

After a detailed analysis of the different possibilities for representing multilinguality at the meta-data level of a Knowledge Based Representation (cf. section 14.1), we have concluded that the 1st option presented in **section 14.1** would better meet NeOn needs. That 1st option exemplified in Figure 39 implied the inclusion of metadata about linguistic information by means of an **extension of the current OMV Core**, what we have called **LexOMV**.

Our aim was to capture the linguistic information included in the Multilingual Ontology Meta-model proposed for NeOn, i.e., that the different Ontology Elements have associated certain linguistic information in different languages. In order to be able to capture such an amount of data at the metadata level we needed to extend the OMV Core and discard the 2nd option presented in section 14.1 because it did not allow us to express as much information as the first one.

As can be seen in Figure 48, we create a new class called `OntologyElement` that allows statements about the different elements to be included in an ontology separately. Since ontologies in NeOn follow the DL paradigm, the different ontology elements will be classes, properties, individuals, etc. However, it is important to note that this model enables the description of ontologies that follow other paradigms. Then, we define a class called `LinguisticElement` in which we include the attributes *name* -referring to the name of the linguistic element: label, definition or source, for example-, and *description* -including an explanation of what is understood under label, definition or source. As it is expected, we also define a class called `NaturalLanguage` with attributes *name*, *description* and *ISOcode* that allows us to refer to the different languages as defined by the ISO standard 639. Finally, we define the class `LinguisticData` in order to associate the multilingual information with the rest of the ontology metadata. So, to express that the piece of linguistic data in question (let us say, definition) is expressed in three languages (e.g. English, Spanish and French) for a certain type of ontology element (e.g. `Class`) in a given ontology, we link the ontology (described in the OMV Core) via the `hasAssociated` relation to the `LinguisticData` class where we integrate all the necessary information using the `hasOntologyElement` property to relate the `Class` ontology element, `hasLinguisticElement` property to relate the *Definition* linguistic element and `isExpressedIn` to relate the *English*, *Spanish* and *French* languages. Furthermore, our extension allows us to describe who the authors and contributors of that linguistic data were by relating the `LinguisticData` class to the `Party` class of the OMV Core.

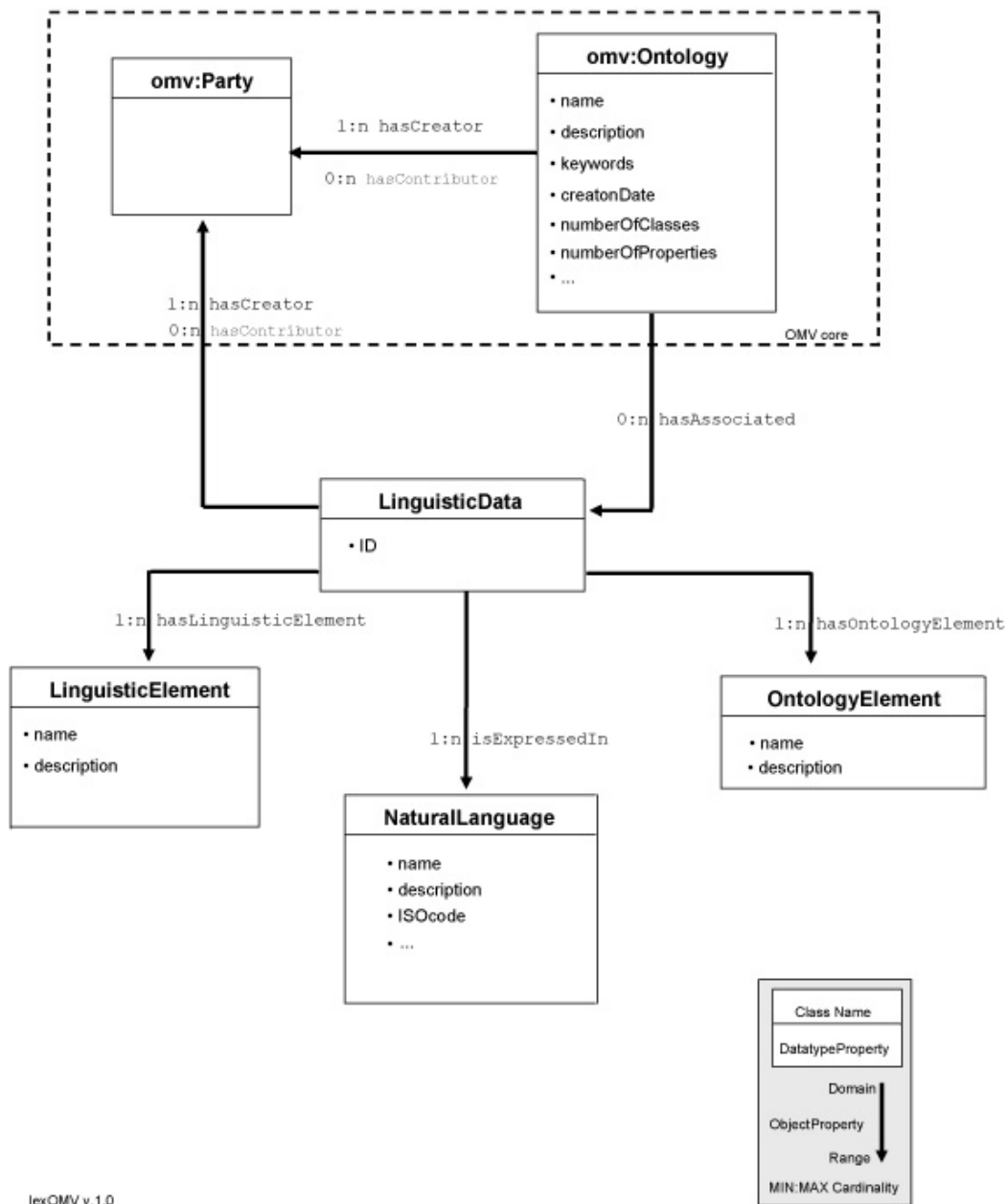


Figure 48: Extension of the OMV Core to capture multilingual data: LexOMV

Relevant bibliographic References:

Aguado de Cea, G. E. Montiel-Ponsoda, and J.A. Ramos Gargantilla. (2007). Multilingualidad en una aplicación basada en conocimiento. TIMM SEPLN, volume 38, pp. 77-97.

Buitelaar, P. Sintek, and M., Kiesel, M. (2006) "A Lexicon Model for Multilingual/Multimedia Ontologies" In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro.

Cabr , M. Teresa; Bach, C.; Estop , R.; Feliu, J.; Mart nez, G.; Vivaldi, J. (2004). "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities". *LREC 2004 Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Languages Resources Association. pp. 87-90.

Hartmann J., Palma R. (2006). *OMV - Ontology Metadata Vocabulary for the Semantic Web, (2006). v. 2.0*, available at <http://omv.ontoware.org/>.

Holi, M. and E. Hyv nen. (2004). "Probabilistic information retrieval based on conceptual overlap in semantic web ontologies". In Proc. Finnish Artificial Intelligence Conference (FAIS'04), vol. 2, Finland.

Miller, G., Beckwith, R., Fellbaum, C. Gross, D. and Miller, K.J. (1990) "Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography*, Vol 3, No.4, 235-244.

Saloni, Z., S. Szpakowicz, and M. Swidzinski. (1990). "The Design of a Universal Basic Dictionary of Contemporary Polish", *International Journal of Lexicography* Vol. 3 no. 1, 1990 Oxford University Press.

Vossen, P. (2004). "EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index". *Semi-special issue on multilingual databases (IJL 17/2, June 2004)*.

17. 1st Prototype of the NeOn Multilingual Ontology Meta-model

This section describes the features and design aspects of the 1st prototype of the *Neon Multilingual Meta-model*, which offers the users a set of functionalities for linguistically enriching the labels of an ontology.

17.1 Requirements specification

The goal of the NeOn Multilingual Meta-model is to provide linguistic information to the different elements that compose an ontology (classes, properties, etc.). Thus, we have developed a plugin, which provides support for managing the linguistic information reflected in the model. This plug-in is based on the *ontology label translation supporting tool LabelTranslator*, fully explained in section 10.2 of this Deliverable. LabelTranslator has been enhanced and its functionalities widened to provide multilinguality to ontologies in NeOn. In the following we summarize the main requirements of the 1st prototype of the NeOn Multilingual Ontology Meta-model.

- LabelTranslator will give support to the translation of ontological labels. In this sense, a label can represent a class name, a property name, etc.
- Linguistic information to be considered (i.e. that LabelTranslator will manage) will be:
 - Label
 - Gloss or Definition
 - Context or Additional Notes (i.e. explanations)
 - Source of knowledge
- LabelTranslator is not meant to update the ontology multilingual information (unless this is ordered by the user), but it will prepare the linguistic information to be updated by the proper agents (export).
- The user will select from the ontology the label to translate or edit. Then, the user will be able to decide whether to translate the label himself or with LabelTranslator's help.
- User interaction with LabelTranslator will follow this schema:
 - The user selects the label to translate
 - The user selects the target language
 - LabelTranslator will look for the relevant information in the lexical resources that have been implemented.

- EWN databases
- Web Resources
 - GoogleTranslate
 - Wiktionary
 - IATE
 - BabelFish
 - FreeTranslation
- LabelTranslator will show the results.
- The information that LabelTranslator will show is:
 - Source of information
 - Translated labels
 - Definition, Context or Notes (if possible).
- If LabelTranslator did not find any information, it would show an information message.
- If LabelTranslator had an error, it would show an error message.

Vocabulary

- Edition: adding linguistic information to an ontology label in the source language.
- Error: a logical / physical error in the system or a system failure (i.e. you have already added some linguistic information to the term, and a connection error or an internal system error occurs when accessing the database).
- Export: prepare the linguistic information for its commitment in the system.
- Incidence: something that happens in the system and is relevant for the user because it does not respond to the normal flow of the application (e.g. error, no results found).
- Ontology Label: any part of the ontology that could be translated / edited.
- Translation: the result of a localization process (i.e. providing a lexical equivalent in a target language).

17.2 NeOn Multilingual Meta-model implementation proposal

In this section we describe the high level architecture that features the 1st prototype and we discuss some of the innovations planned for the 2nd prototype of the *Neon Multilingual Meta-model*. We propose a three layer approach in order to implement the multilingual model. The multilingual information requirements identified in section 12.3 have been taken into consideration in the proposed architecture.

In Figure 49, the architecture of the 1st prototype is shown. The first layer encapsulates the graphical user interfaces that permit the interaction with the user. The GUIs implemented in this layer allow a semi-automatic translation of a specific ontology label. The *LabelTranslatorService* implements the business functionality of the multilingual model (in the second layer). This service encapsulates some functionalities such as: translation of ontological labels, ranking of the senses of a translated label, etc. Finally, the third layer provides the repository in order to store the linguistic information. The current NeOn Toolkit views are used for storing the multilingual information.

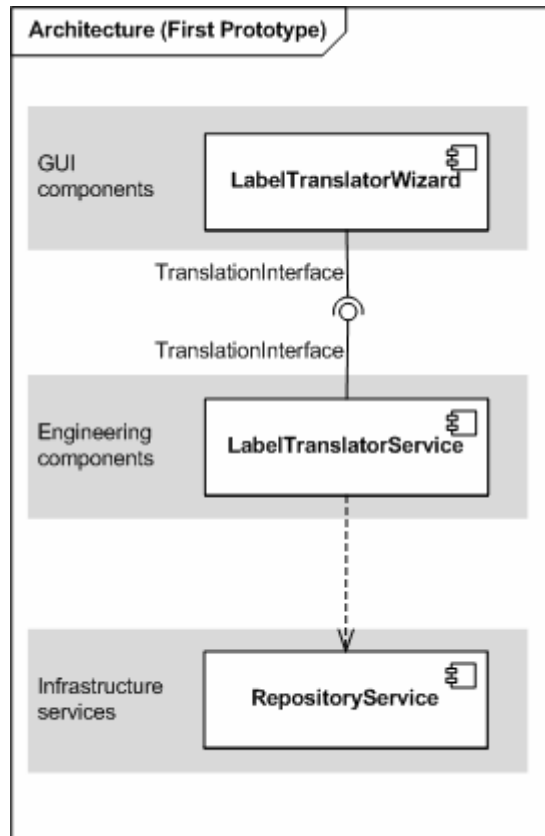


Figure 49: Three layer architecture of the 1st prototype of the NeOn Multilingual Meta-model

The architecture above described covers only the basic functionalities of the requirements for multilingual information representation included in the different NeOn WPs (see section 12.3). Thus, a 2nd prototype is planned in order to fulfill all requirements. Figure 50 shows how the current prototype can be enriched.

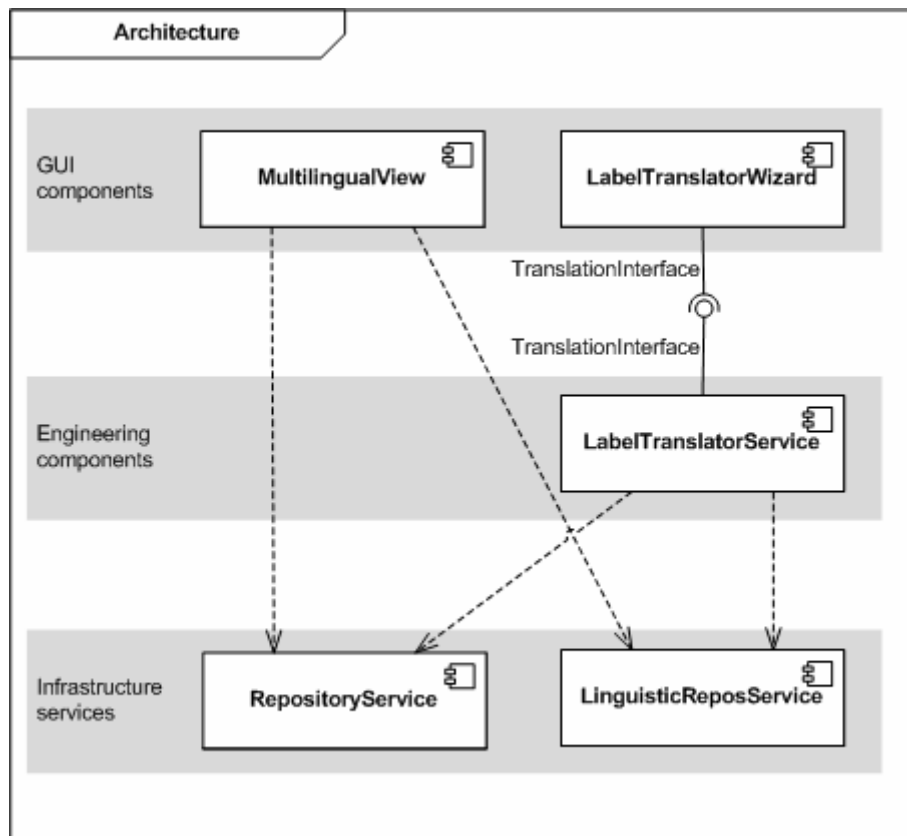


Figure 50: Schema of the 2nd prototype of the NeOn Multilingual Meta-model

As can be seen in Figure 50, a *MultilingualView* will be added which would contain a set of GUIs for editing the multilingual information. A new repository (*LinguisticReposService* in the figure) will be also used in order to store the linguistic information. Consequently, the linguistic information will be stored in two places at the same time.

In the next sections we describe in more detail the functionalities that characterize the current 1st prototype.

17.3 Description of the 1st Prototype of the NeOn Multilingual Meta-model

Currently, the possibility of adding multilingual information to ontologies is not yet very sophisticated so as to access information in a seamless and transparent way. The following problems have to be solved in order to enable users to access multilingual information:

1. translating words,
2. disambiguating word senses, and
3. presenting the multilingual results appropriately.

Here, we consider these problems and describe the architecture proposed for enriching an ontology with linguistic information.

17.3.1 Architecture

First, we introduce the *LabelTranslator plugin* which extends the NeOn toolkit (an extensible Ontology Engineering Environment) for supporting the translation of ontological labels using relevant information obtained from different lexical resources. We describe here the functionalities that characterize the current 1st prototype. The main components of the *LabelTranslator plugin* are shown in Figure 51.

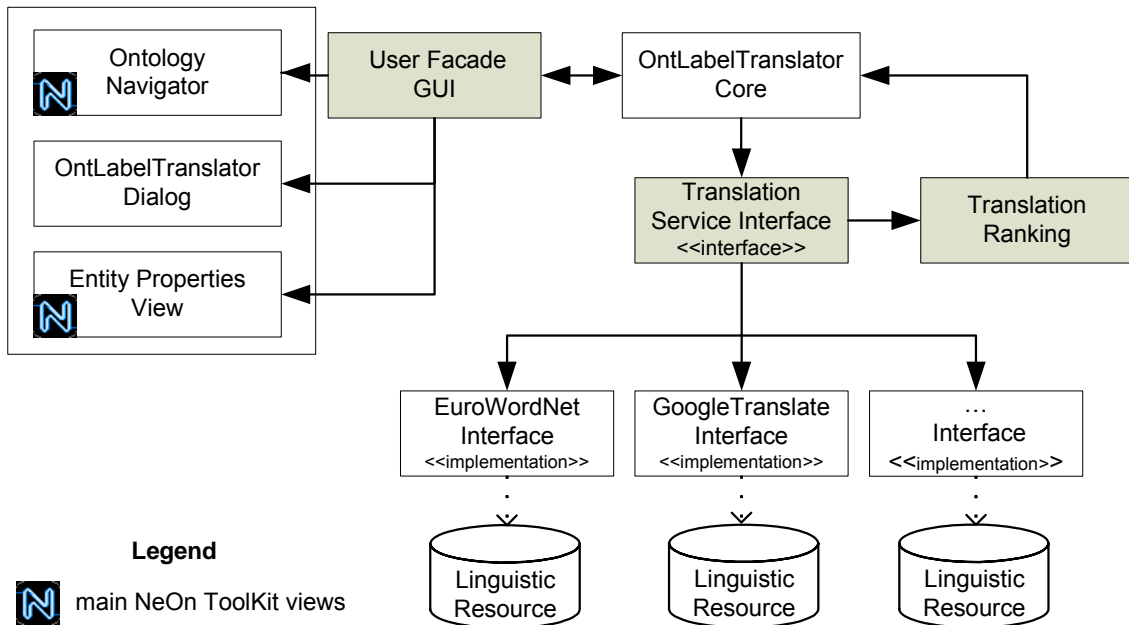


Figure 51: Main components of the *LabelTranslator* plugin

User Interface

The *User Facade* controls the GUIs in order to show the multilingual results appropriately. *LabelTranslator* provides additional extension-points to modify the main components of the NeOn Toolkit, the *Ontology Navigator* and the *Entity Properties View*. The *Ontology Navigator* is a completely modifiable and extensible view on (not necessarily) ontology elements, and it offers a perspective over ontological data in the NeOn Toolkit-style. By right clicking on a frame (classes or properties, for example), a typical contextual menu appears. In order to support the translation of ontological labels, the *LabelTranslator* plugin provides a further *extension* to current actions of the contextual menu: the *Translate* action. In Figure 52, we present a screenshot of the *Ontology Navigator View* with the extension to translate an ontology label.

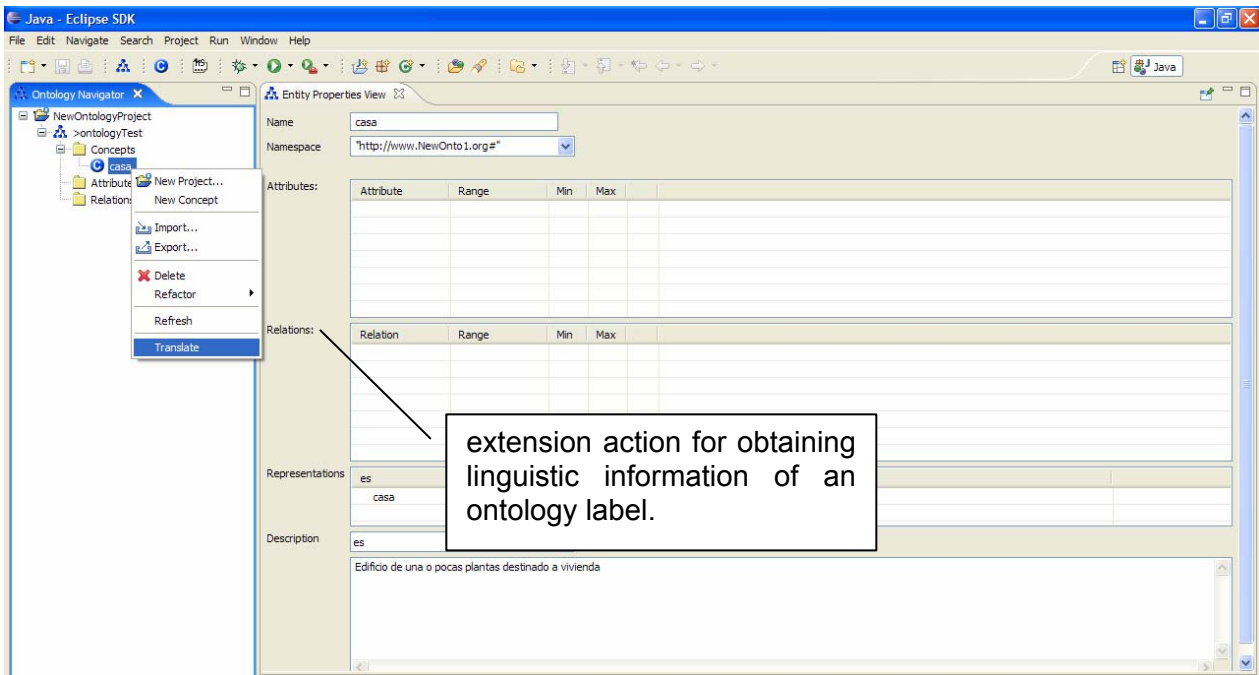


Figure 52: A screenshot of the Ontology Navigator with the action “Translate”

Another view used as user interface for the LabelTranslator plugin is the Entity Properties View, which shows property pages for the elements in the user interface. In this case the plugin does not add extensions; however, some fields and tables (that show linguistic information) are filled in runtime, according to the modalities decided by the Label Translator core. In Figure 53, we show a screenshot of the Entity Properties View with some linguistic information updated from the results obtained by LabelTranslator.

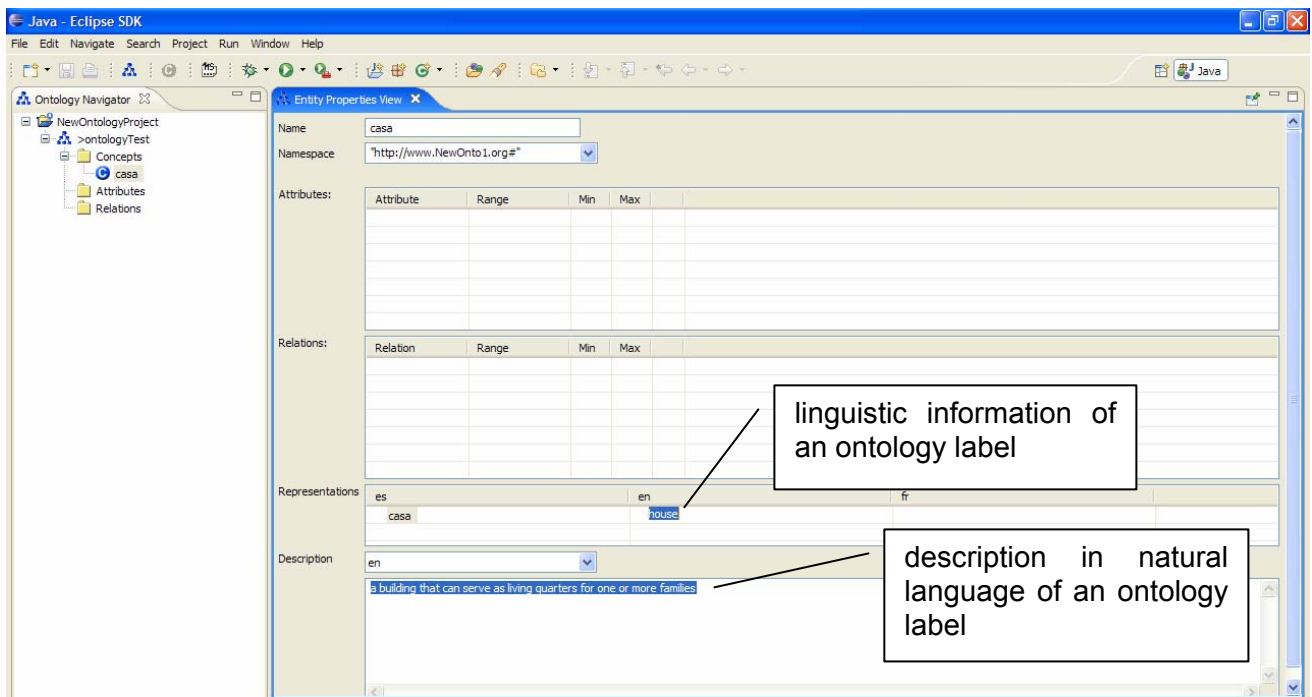


Figure 53: A sample of linguistic information in the Entity Properties View

In the 1st prototype, a dialog that shows the candidate translations of the ontology label under consideration has been additionally designed. Figure 54 presents a screenshot of the LabelTranslator dialog with the results obtained after the translation process of an ontology label.

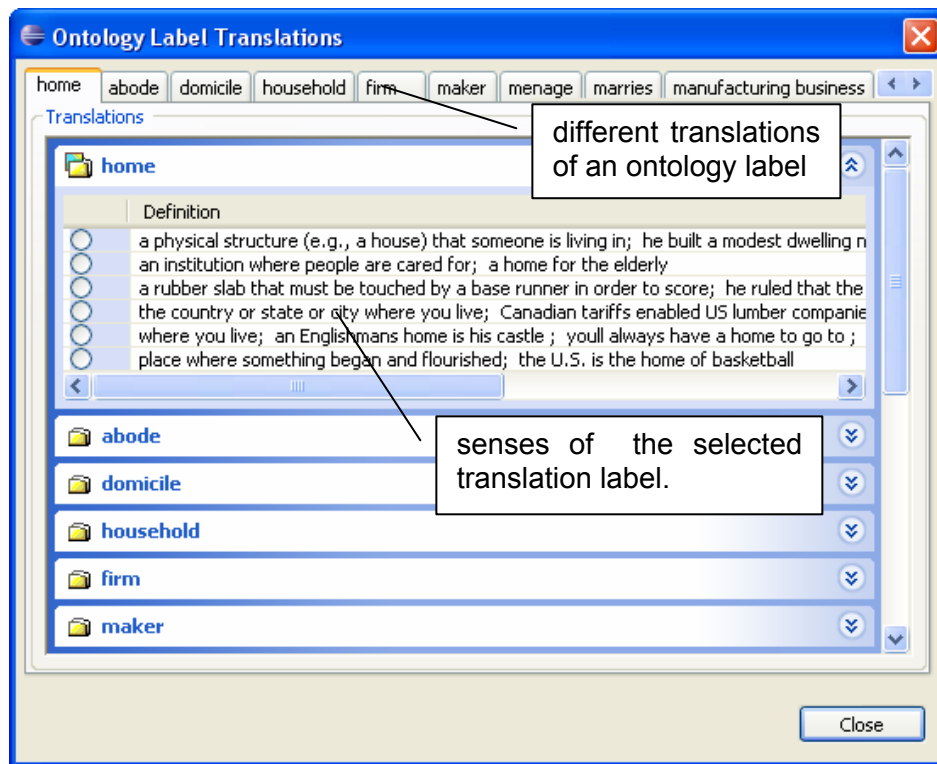


Figure 54: User dialog with the translation results of an ontology label

Translation service features

In order to automatically extract translations of an ontology label, we use a translation service, which relies on different linguistic resources. A linguistic resource contains sets of language data and descriptions, that can be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems. In particular, we include lexical databases, bilingual dictionaries and terminologies as linguistic resources.

This service has been implemented as a java package on its own, which can externally be imported by any application willing to exploit natural language resources such as lexicons and terminological dictionaries. We have developed several implementations of the Translation Service interface for: EuroWordNet, IATE, GoogleTranslate, Wiktionary, Babelfish and FreeTranslation. All interfaces provide translations for several languages and return a flat list of linguistic expressions as result. Additionally, services include so-called “glosses” (if they exist) offering a short definition of the term under consideration in both source and target language.

The translation service method uses also a compositional method in order to perform the translations of compound words. Compound words are often not contained in linguistic ontologies such as EuroWordNet. However, the meaning of such a compound word can be obtained in many cases from the combination of the meanings of the different words that form the compound word. If a person, for example, does not know the meaning of a compound word he/she tries to decompose it in order to extract the sense of each component. In order to understand the word sense as a whole, people frequently try to translate the individual word parts in their own language and then try to understand the linguistic context. The compositional method relies on a translation-candidate collection and translation selection. The compositional method first searches for translation candidates of a given compound word and then finds the translations for the candidates.

Finally, if the term label is not found, the user may enter his/her own translation (together with the definition).

Translation Ranking Method

The *translation ranking method* sorts the translations of an ontology label into sense lists based on contexts. Because many translations contain ambiguities, putting the correct translations on the top of the result list saves users time in consultation. Given an ontology label to be translated, the main steps of the algorithm are:

1. The method, in parallel:
 - a. determines the context of the selected label in the ontology extracting the adjacent labels.
 - b. obtains the translations of the selected label using the translation service.
2. A vector representation is created for the senses of each translation (using the source language) and adjacent label (that composes the context of the label under consideration).
3. A disambiguation method⁷³ is used for disambiguating the possible set of senses generated by the translation process. This is carried out by comparing the senses associated to translation and label context entries.
 - a. Senses of each translation are sorted according to similarity with the context of the translated ontology label (determined in the first step).
4. Using a multilingual resource, the method obtains the senses in the specified target language.
5. The method shows the sorted results to the user.

In the next section we present a more detailed description of the steps performed by the translation ranking method.

17.3.2 Ranking for ordering translations

The method takes as input an ontology label in a specific source language and returns a sorted list of linguistic information (according to similarity with the context of the selected ontology label) in a target language. The proposed method is outlined in Figure 55.

⁷³ Word sense disambiguation is the process of assigning a meaning to a particular word based on the context in which it occurs

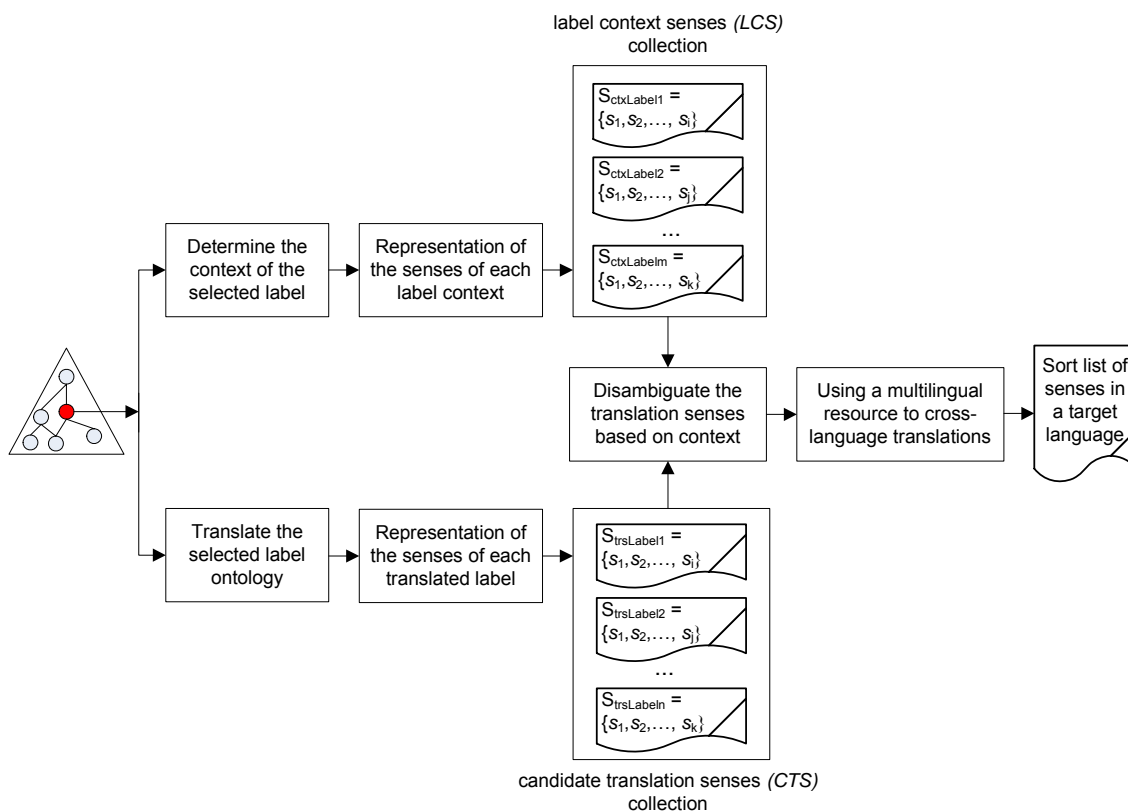


Figure 55: Main steps of the translation ranking method

Determining the context of the ontology label

Contexts have the generic property of disambiguating the lexical meaning of a word. For example, the term *bank* has a different meaning in a *geographical* context than in a *financial* context and therefore evokes different concepts. In order to determine the context of an ontology label (*LabelContext*), we retrieve the set of labels associated with the label under consideration.

LabelContext comprises a set of names, which can be: direct label names and/or attributes label names, depending on the type of term (identified by the label) that is being translated. For instance, a *class* label can have as context a combination of two aspects: the labels in the hierarchy which are adjacent to it (both hypernyms and hyponyms labels), and its attribute labels. The context of a *property* label can be represented by the labels that represent the domain and range and the adjacent⁷⁴ hierarchy labels. The experiments section presents a study of the influence of each of these context combinations on the disambiguation accuracy.

Representation of the senses of ontological labels and translations

⁷⁴ Properties are organized in hierarchies in some ontologies.

Traditional information retrieval typically represents data using a bag of words while data mining typically uses a highly structured database representation. In order to represent the senses of both ontological labels (those that describe the context of the ontology label under consideration) and the labels extracted from the translation process, we propose a middle ground method between bag-of-words document retrieval and highly-structured data mining. The idea of using a set of words to express the semantics of a concept is inspired on the approach taken in WordNet. We create a document-style representation of each entity based on the following definition.

Definition. The representation of the senses of both ontological and translation labels is a set of names that describe the sense of an ontology or translation label in a determined context. Keywords and/or additional phrases (explanatory glosses) constitute the elements of the sense representation.

We use the notation $ctxSense(ctxLabel)$ and $trsSense(trsLabel)$ to denote the lexical representation of the sense of an ontology label or the sense of a translation label respectively.

$$ctxSense (ctxLabel) = \{ \{ k_1, k_2, \dots k_n \}; [gloss] \} \text{ and}$$

$$trsSense (trsLabel) = \{ gloss \}$$

Here, $\{k_1, k_2, \dots k_n\}$ and $\{gloss\}$ are the elements of both $ctxSense(ctxLabel)$ and $trsSense(trsLabel)$. k_i represents the labels extracted (if applicable) from the semantic relations found in a formal model (for example an ontology) as synonyms, hypernyms, hyponyms, and meronyms. *Gloss* represents the main words that describe the term in natural language. Main words are referred to as those words which are not filtered out prior to the processing of natural language data. These filtered words are known as ‘stop words’. Please, note that the set of keywords describing the gloss of $ctxSense$ is optional.

For example, the sense representation of the ontology label “java”, when used in the “computer science” context, could be,

$$ctxSense (java) = \{ \{ \text{“java”, “programming language”}; \{ \text{a simple platform independent...} \} \}$$

If used in the context of “travelling” could be,

$$ctxSense' (java) = \{ \{ \text{“java”, “island”, “vacation_destination”}; \{ \text{a island of Indonesia..} \} \}$$

Note that to illustrate our example the “stop words” in the gloss of both samples have not been deleted. This specification allows a machine to retrieve, compare, etc. concepts or classes in an ontology. These unique combinations of keywords describe the vocabulary used to model the senses of an ontology label and/or a translation label.

Disambiguating the senses of the translations.

The main problem to be solved in order to enable users to access multilingual ontology information is the disambiguation of a word. This refers to the fact that there can be more than one possible entry in the lexical resource that relates to the label in the ontology (ambiguity problem). One example is given by the word *book*. It has different meanings (a written work that has been published, arrange for and reserve in advance, etc) that can be recognized by means of the context.

Here we present a method of word sense disambiguation that assigns a sense to a target word by maximizing the relatedness between the target and its neighbours. We carry out disambiguation in relation to the senses retrieved from the linguistic resources described in the previous section. We use both its semantic relation structure (if possible) and glosses of word meanings to measure semantic relatedness. The method is not supervised, and does not require any manually created sense-tagged training examples. The underlying presumption of this disambiguation method is that words that occur together in a sentence should be related to some degree.

In the following we describe the method: let us suppose that the ontology label to be disambiguated has the name *label*, and after executing the translation process it has yielded n translations: $T = \{\text{trsLabel}_1, \text{trsLabel}_2, \dots, \text{trsLabel}_n\}$. For each translation label the plugin retrieves its corresponding senses:

$$S_{\text{trsLabel}_1} = \{s_1, s_2, \dots, s_i\}; S_{\text{trsLabel}_2} = \{s_1, s_2, \dots, s_j\}; \dots S_{\text{trsLabel}_n} = \{s_1, s_2, \dots, s_k\}$$

where S_{trsLabel_i} represents the set of senses of the i^{th} translation label. In order to represent the senses of each translation, the definition described in the previous section is used. So, a candidate translation sense (CTS) collection is obtained.

$$\text{CTS} = \{\text{trsSense}(s_1^{\text{trsLabel}_1}), \text{trsSense}(s_2^{\text{trsLabel}_1}), \dots, \text{trsSense}(s_1^{\text{trsLabel}_2}), \text{trsSense}(s_2^{\text{trsLabel}_2}), \dots, \text{trsLabel}(s_1^{\text{trans}_n}), \text{trsLabel}(s_i^{\text{trans}_n}), \dots\}$$

where $\text{trsSense}(s_j^{\text{trsLabel}_k})$ is a vector with the elements extracted of the words that compose the j^{th} sense corresponding to k^{th} translated label trsLabel .

Now, let us suppose that the context (*LabelContext*) of *label* comprises several names: $\text{LabelContext} = \{\text{ctxLabel}_1, \text{ctxLabel}_2, \dots, \text{ctxLabel}_m\}$, which depend on the type of term (associated to label) that is being translated. Each of these context names has a list of corresponding senses, for instance ctxLabel_j has p senses: $S_{\text{ctxLabel}_j} = \{s_1, s_2, \dots, s_p\}$. The linguistic information of the senses of each context label is represented using the same definition used for the senses of each translated label. In this way a label context sense (LCS) collection is obtained.

$$\text{LCS} = \{\text{ctxSense}(s_1^{\text{ctxLabel}_1}), \text{ctxSense}(s_2^{\text{ctxLabel}_1}), \dots, \text{ctxSense}(s_1^{\text{ctxLabel}_j}), \dots, \text{ctxSense}(s_p^{\text{ctxLabel}_j})\}$$

where $\text{ctxSense}(s_j^{\text{ctxLabel}_k})$ is a vector that represents the j^{th} sense of k^{th} context label ctxLabel . The elements of the vector are extracted from the semantic relations of ctxLabel_k in the ontology plus the words that constitute the gloss of $s_j^{\text{ctxLabel}_k}$.

The chosen sense for *label* is given by:

$$\text{Sense (label)} = \text{Max} (\text{SenseScore} (\text{trsSense}_i, \text{LCS}))$$

where i is the i^{th} representation of the senses of each translated label in *CTS* (i goes from 1 to n). The chosen sense is the one with the greater value of *SenseScore*, which is given by:

$$\text{SenseScore} (\text{trsSense}_i, \text{LCS}) = \sum_{j=1}^m \text{Similarity} (\text{trsSense}_i, \text{ctxSense}_j)$$

where m is the number of elements in the vector that represent the senses of the context labels (LCS). The *SenseScore* is the sum of the similarity between each one of the vectors of *TCS* and *LCS*.

In order to compute the similarity between the senses of each context and the translated label, we apply an adapted version of the Lesk algorithm (Banerjee and Pedersen 2003), which uses the gloss of a word and an overlap scoring mechanism. In particular, the method uses not only the gloss/definition of the sense but it also considers the meaning of related words in order to compute the total score between two senses. The translated labels are sorted according to the similarity of their senses with the context of the selected ontology label.

Using a multilingual resource to cross-language translation

In order to discover the correspondences in the target language of the sorted list of linguistic data mentioned in the previous steps, we use a multilingual resource. In particular for the 1st prototype we rely on EuroWordNet (Vossen 1997), which provides a list of word senses for each word, organized into synonym sets (SynSets). The multilingual retrieval of a word sense (SynSet) is done by using the ILI entries (explained in more detail in section 9.1). For example, when a synset, e.g. “bank” with the meaning “financial institution”, is retrieved in the English WordNet⁷⁵, its SynSet-ID can be used to retrieve the same concept in all other language-dependent WordNets (German, Spanish, etc.) that describe the same concept with the same ID, but naturally contain the word description in its specific language.

17.4 Use Cases of the 1st Prototype.

This section describes some use cases of the user interaction with the 1st prototype of the NeOn Multilingual Meta-model.

17.4.1 Use Case: Add Language

Overview

The ontology editor needs to create a new language to work with the ontologies. This language is used by the different views of the system to show the corresponding information.

⁷⁵ [http:// wordnet.princeton.edu/](http://wordnet.princeton.edu/)

GUI Prototype

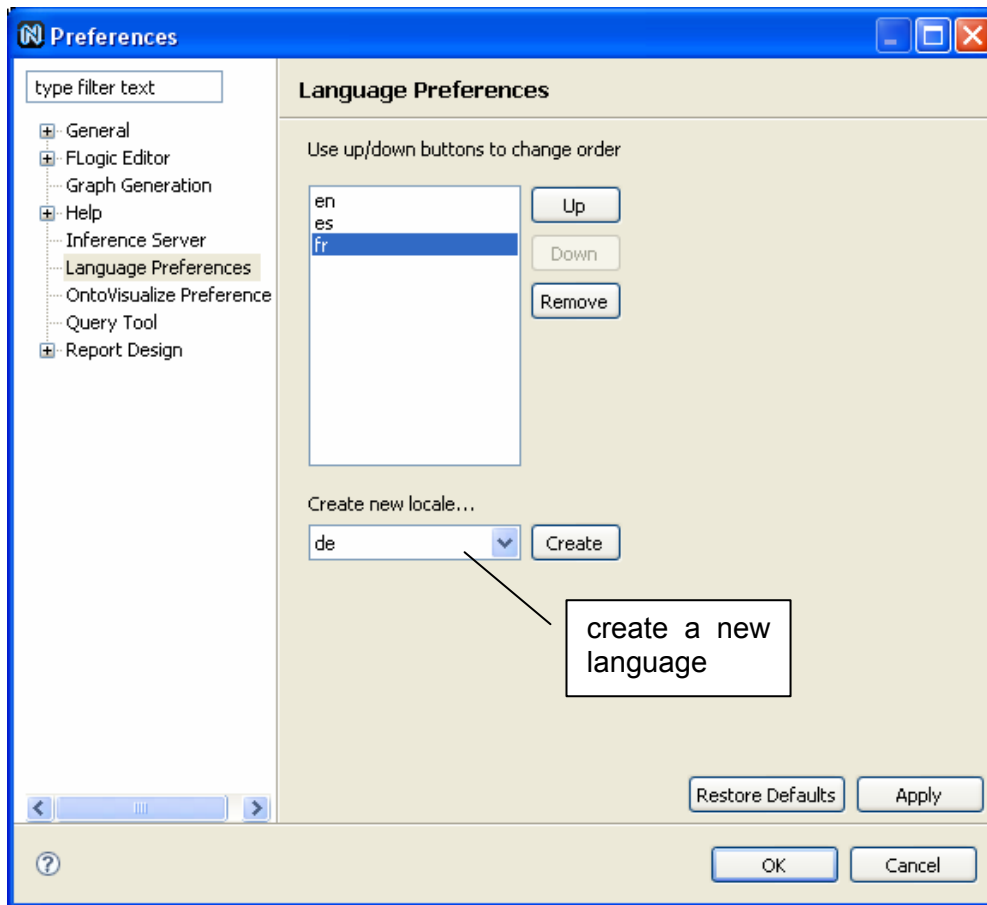
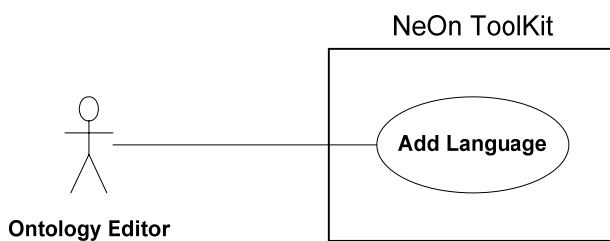


Figure 56: GUI prototype of the language preferences

Detailed description



Primary Actor: Ontology Editor

Stakeholders and Interest

The Ontology Editor wants to incorporate a new language for an entire ontology.

Preconditions: the Ontology Editor has been logged in and has the permissions for adding a new language.

Success Guarantee: a new language is created

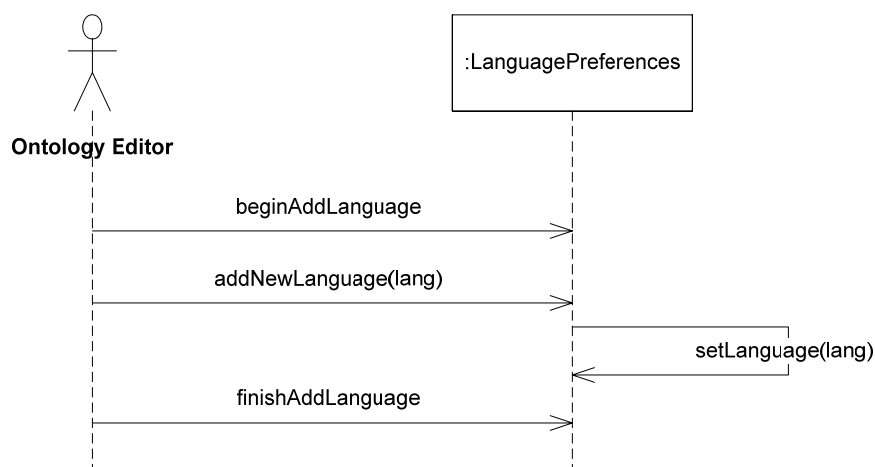
Fail Guarantee: The System remains as it was before the execution of the use case.

Extension Points:

Main Success Scenario (or Basic Flow):

1. The Ontology Editor informs the System that he wants to add a new language.
2. The Ontology Editor creates a new language.
3. The System creates the new language and adds it to the list of language preferences.

Frequency of Occurrence: medium

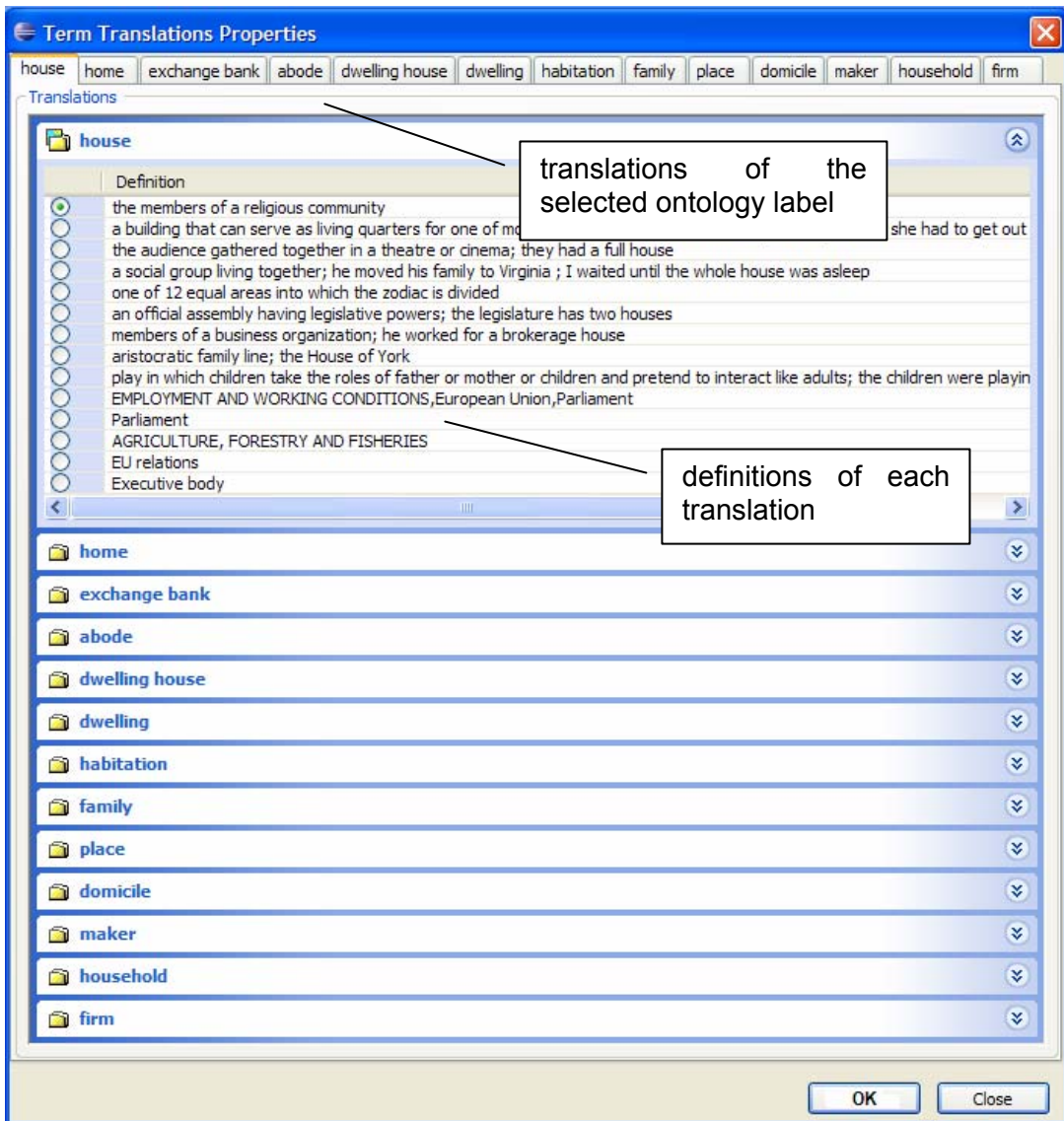
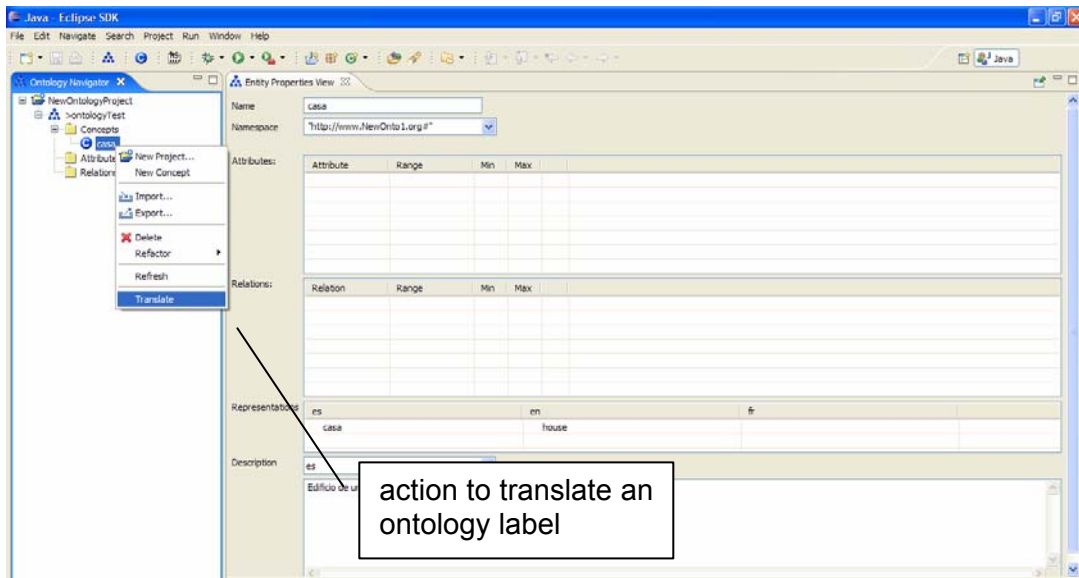


17.4.2 Use Case: Translate an ontology label

Overview

The Ontology Editor wants to translate an ontology label in a target language using LabelTranslator plugin. Additionally, the user can manually edit or delete any part of the linguistic information of the selected term.

GUI Prototype



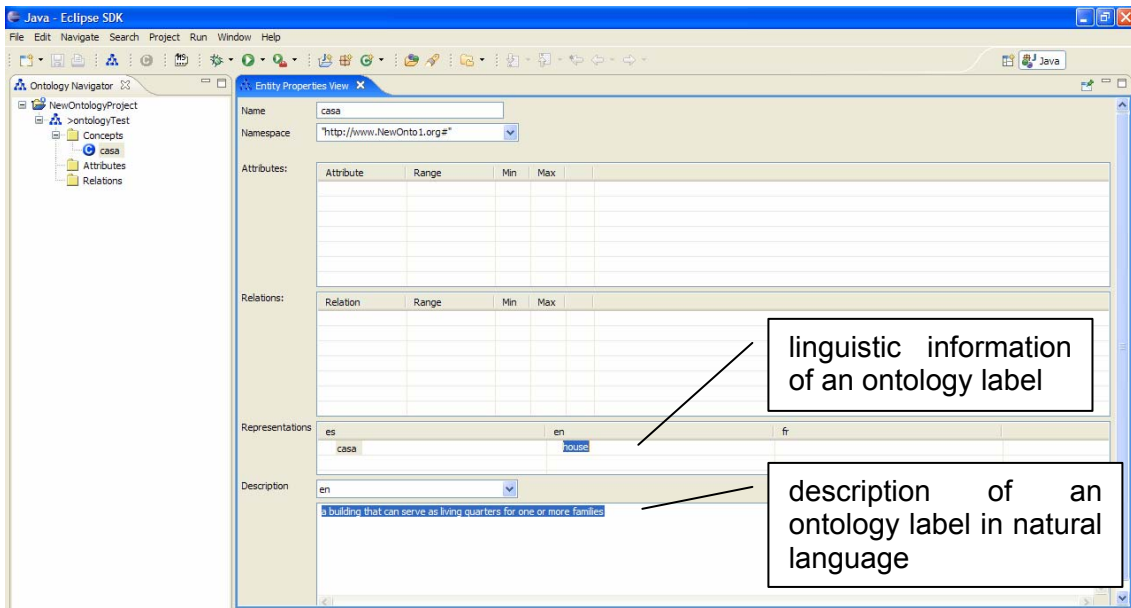
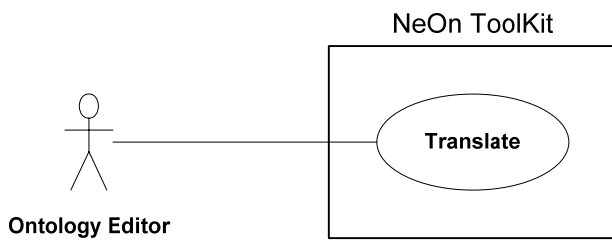


Figure 57: GUIs to translate an ontology label using LabelTranslator

Detailed description



Primary Actor: Ontology Editor

Stakeholders and Interest:

The Ontology Editor wants the LabelTranslator plugin to help him/her with the translation of an ontology label.

Preconditions:

The Ontology Editor is working with an ontology
The Ontology is open

Success Guarantee: LabelTranslator will introduce the results at the proper place

Fail Guarantee:

LabelTranslator will return an error message.
No linguistic information will be returned

Main Success Scenario (or Basic Flow):



1. The user is working with an ontology
2. The user selects the ontology label to be translated into other target languages using LabelTranslator.
3. The user selects the target language.
4. LabelTranslator will look for the relevant information in the corresponding lexical resources.
 - a. EWN MySQL databases
 - b. Babelfish
 - c. GoogleTranslate
 - d. IATE
 - e. Wiktionary
 - f. FreeTranslation
5. LabelTranslator ranks the results (linguistic information) and proposes a label to the user
6. LabelTranslator will put the results at the proper fields.

Extensions (or Alternative Flows):

- 5a. The user chooses a different label instead of the proposed label by LabelTranslator.
 1. The use case continues on step 6.

Related Use Cases:

Edit multilingual labels

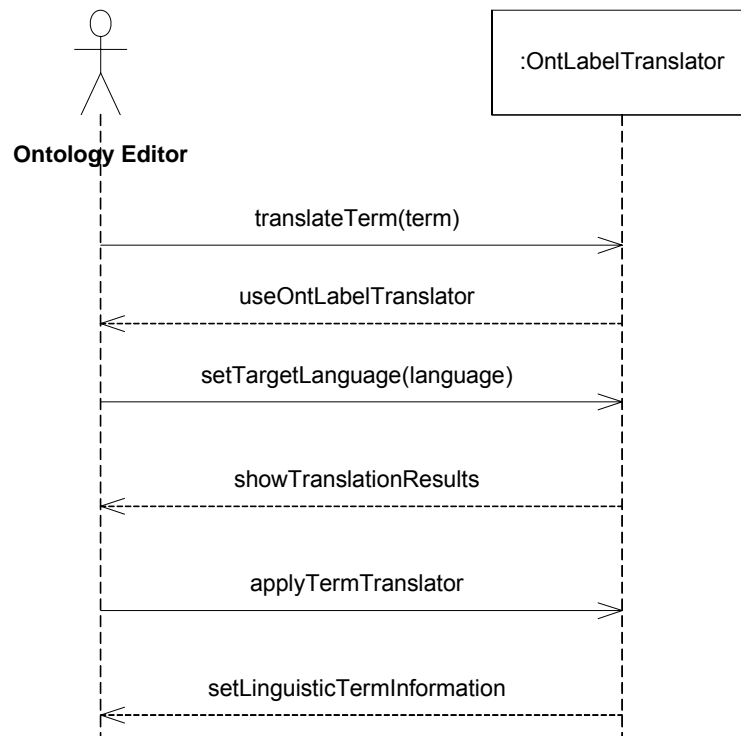
The user is working with an ontology

The user edits the multilingual information and adds the results at the proper fields

Delete multilingual labels

The user is working with an ontology

The user deletes the multilingual information.



Relevant bibliographic references

Banerjee, S and T. Pedersen. (2003). Extended gloss overlaps as a measure of semantic relatedness.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, Zurich.

18. Future work

As already introduced in section 17.2 and represented by the schema in Figure 50, the 2nd Prototype of the NeOn Multilingual Meta-model will consist of a Linguistic Information Repository (LIR) linked to the Ontology Metamodel. The LIR has been presented and fully described in section 16.1. This is the Multilingual Ontology Meta-model proposed for NeOn after carrying out an extensive survey of multilingual resources and taking into consideration NeOn WPs requirements.

For the next phase of work, then, we foresee the following two main tasks, which primarily form an extension of the work already undertaken.

1. Adjustment of the LIR model according to emergent user needs and new developments in standardization initiatives, especially MLIF, in which the authors are actively engaged.
2. Implementation of the 2nd Prototype of the Multilingual Ontology Meta-model.
3. Further development of LabelTranslator

Appendix 1

Order	ACRONYM	URL	Domain	Level
1	EUROVOC	http://europa.eu.int/celex/eurovoc/cgi/sga_doc?eurovoc_dif!SERVEUR/menu!prod!MENU&langue=EN		General
2	UNESCO	http://databases.unesco.org/thesaurus/		General
3	UNBIS	http://unhq-appspub-01.un.org/LIB/DHLUNBISThesaurus.nsf		General
4	OECD	http://info.uibk.ac.at/info/oece-macroth/en/508.html	ECONOMIC CO-OPERATION AND DEVELOPMENT	General
5	SOSIG	http://sosiq.ac.uk/roads/cgi-bin/thesaurus.pl	Social Science	General
6	GETTY	http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html	Geographical names	General
7	Astronomy Thesauru	http://msowww.anu.edu.au/library/thesaurus/english/	Astronomy	General
8	WORDNET	http://www.cogsci.princeton.edu/cgi-bin/webwn2.0?stage=2&word=limnology&posnumber=1&searchtypenumber=2&senses=&showglosses=1		Specific
9	NBII	http://thesaurus.nbii.gov/	Biology	Specific
10	NAL	http://agclass.nal.usda.gov/agt/search.htm	Agriculture	Specific
11	USAID	http://www.dec.org/pdf_docs/PNACD400.pdf	development	Specific
12	GEMET	http://www.eionet.eu.int/gemet	Environment	Specific
12	GEMET	http://eea.eionet.eu.int:8980/irc/Download/kjedA-JCmmGDso6e5BXEwozPOfDYu3Gi-oCYw6OlawErHX45Z1iH4pYxtvF37-3HY/GemAlph.pdf	Environment	Specific
13	CABI	http://www.cabi.org/DocServer/default.aspx?id=11111	Agriculture	Specific
13	CABI	http://library.vetmed.fu-berlin.de/cab/Thesaurus.html	Agriculture	Specific
13	CABI online	http://194.203.77.66/Search.asp	Agriculture	Specific
14	ASFA	http://uk2.csa.com/helpV3/ab.html	Fisheries	Specific
14	ASFA	http://uk2.csa.com/htbin/ccfdisp.cgi?fn=wais/data/thes/asfithes.ccf&st=A&fmt=5&ldtag=TR	Fisheries	Specific
15	CHEMISTRY GENERAL	http://antoine.frostburg.edu/chem/senese/101/glossary.shtml	chemistry	Specific
16	BANANA	http://www.inibap.org/bdd/thesaurus_EN.htm	Plants	Specific
17	ADM	http://www.med.univ-rennes1.fr/htbin/adm/reponse.pl?menu=menu.html	Medicine (FR)	Specific
18	MESH	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=MeSH&term=	Medicine	Specific
19	ILO	http://www-ilo-mirror.cornell.edu/public/english/support/lib/dblist.htm		
21	HURIDOCS	http://www.huridocs.org/mt.htm	Human Rights	Specific
22	AGCOM Thesaurus	http://web.aces.uiuc.edu/agcomdb/thesaurus.html	Agricultural communications	
23	National Digital Archive of Datasets	http://www.ndad.nationalarchives.gov.uk/search/thesaurus/terms/list1.htm#Agriculturaleducation		
26	The micro-thesaurus on occupations by ILO categories	http://www.huridocs.org/mt10.htm	Occupations	
27	The AOD Thesaurus. Annotated Hierarchy. field/discipline/occupation	http://etoh.niaaa.nih.gov/AODVol1/aodhns.htm#SL12-10	Occupations	
28	The Dictionary of Agricultural Occupations	http://www.cnr.berkeley.edu/ucce50/ag-labor/7manual/7dao.htm	Occupations	
29	The Texas Farm Bureau website	http://www.txfb.org/AgClass/resource/AITCrq28.htm	Occupations	
30	Wikipedia	http://en.wikipedia.org/wiki/Category:Agricultural_occupations	Occupations	
31	ACRONYMS	http://www.acronyma.com/	General	General
32	tesauro sobre biodiversidad de Colombia	http://www.siac.net.co/sib/tesauros2/WebModuleTesauros/index.jsp	Biodiversity (Spanish)	Specific
33	small-scale food processing equipment	http://www.fao.org/docrep/x5424e/x5424e00.htm	Food Processing	Specific
34	GBIF	http://www.gbif.net/portal/ecat_browser.jsp?termsAccepted=true	Taxonomic names	Specific
35	Animal Diversity Web	http://animaldiversity.ummz.umich.edu/site/accounts/classification/Animalia.html	Taxonomic names	
36	AGRICOLA Thesaurus for Animal Use Alternatives	http://www.nal.usda.gov/awic/alternatives/alfact.htm	Animals	Specific

Appendix 2

The current database is consequently undergoing a revision and will be restructured. The final result will be a concept-based repository called the Concept Server (CS). Relationships between concepts will be made more explicit in order to make better use of them. AGROVOC managers will be created in order to maintain the CS for each language. To promote interoperability, it will be possible to export from the CS in SKOS and OWL formats.

The final workflow is represented as follows in Figure 58:

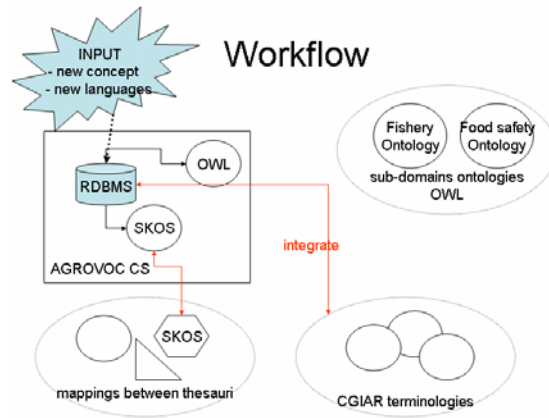


Figure 58: AGROVOC planned revision workflow

Once the above mentioned revisions are done, a new maintenance tool (workbench) will be developed to allow for distributed access to each AGROVOC manager (cf. Figure 59) for maintaining specific languages and/or specific domains.

The overall workflow of the CS management is shown in the following schema:

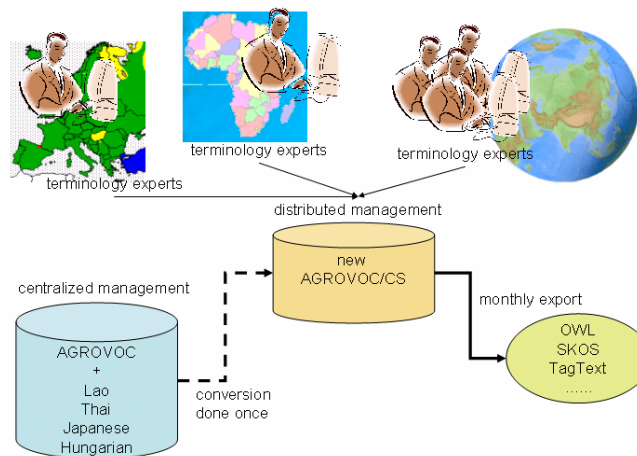


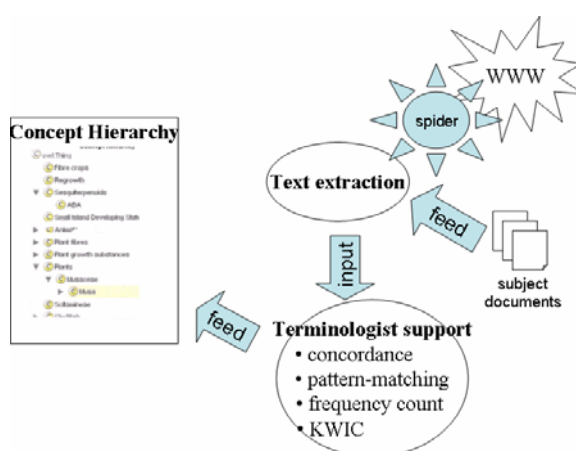
Figure 59: AGROVOC management tool

AGROVOC managers will update the CS by enriching subject coverage and multilingualism, and representing their own “world view”, without starting from a specific language to be translated.

The workbench tool will be structured in such a way that terminologists will have powerful, additional tools for discovering and identifying new concepts, synonyms of existing concepts, etc.

One of these additional tools available for terminologists will be the software **Tropes Zoom**⁷⁶, a Semantic Search Engine and Text Analysis, which already incorporates the AGROVOC Thesaurus. Tropes Zoom has been developed by Semantic-Knowledge, a consortium of several Europe's linguistic software companies.

Tropes Zoom system includes fast Natural Language Information Retrieval system Integrated Web Spider, built-in Semantic Networks and on-the-fly Semantic classifications, among other functionalities. This work has been carried out in partnership with the CIRAD, an Agricultural Research Centre working for international development.



⁷⁶ Tropes Zoom: <http://www.semantic-knowledge.com>