SIERA

Integrating Sina Institute into the European Research Area
FP7- 295006

Deliverable D3.2

Arabization and Multilingual Knowledge Sharing
- Final Report on Research Setup

Mustafa Jarrar (BZU), Adnan Yahya (BZU), Ali Salhi (BZU), Mamoun Abu Helou (BZU), Bassem Sayrafi (BZU), Mahdi Arar (BZU), Jamal Daher (BZU), Amanda Hicks (BBAW), Christiane Fellbaum (BBAW), Stefano Bortoli (UNITN), Paolo Bouquet (UNITN), Rute Costa (UNL), Christophe Roche (UNL), Matteo Palmonari (UNIMIB)

Document Identifier	SIERA FP7-INCO-2011-6 (295006) /2014/D2.3
Version	Version 1.0
Date	26/9/2014
State	Final Version
Distribution	**PP =** Restricted to other programme participants (including the Commission Services)

Deliverable Number	D2.32	Title	Final report on research setup
Work Package Number	2	Title	Networking and Research Setup
Nature	Prototype ⊠ Report Dissemination		
Dissemination Level	Public Consortium ⊠		

Executive Summary

The Work Package 2 (WP2) "Networking and Research Setup" aims to enable close and sustainable scientific cooperation between EU scientists and BZU Sina Institute and to enhance the institute's capacity in scientific cooperation through carrying out several activities:

- Facilitate co-supervision of PhD students (Task 2.1),
- Set up joint research and cooperation (Task 2.2), and
- Encourage co-authoring of scientific articles (Task 2.3),
- Self evaluation (Task 2.4).

This Deliverable is the final report on the consortium's activities carried out in the Task 2.2: "Set up joint research and cooperation". It covers all activities and outcomes of research and cooperation setup, which can be summarized as the following:

Four EU multilingual knowledge portals developed in previous EU FP7 and eTen projects have been Arabized, i.e., extended to support Arabic:

MICHAEL, in section 3, is a thesaurus-oriented portal for the promotion and valorization of digital cultural content. This cultural heritage portal provides a multilingual service to explore digital collections from museums, archives, libraries, etc. Unfortunately, MICHAEL was not full Arabized, as the source code of MICHAEL portal was not available during the project, hence only some related tasks were carried out, such us partial translation of MICHAEL's thesaurus. The project board decided to Arabize the two other portals: OKKAM and Organic.Edunet.

KYOTO, in section 4, is an ontology-oriented wiki-portal in ecology and environment domain. Unlike other portals, KYOTO supports SPARQL queries over structured data and using a rich ontology. Thus, KYOTO ontology has been mapped to the top and core concepts of Arabic Ontology, which is the important step toward integration into the KYOTO architecture, and the potential to facilitate integration with large scale knowledge bases such as DBPedia and GeoNames.

OKKAM, in section 5, is an entity management system, and a spin off by UNITN. OKKAM was Arabized (its Interface, Ontology, content, an searching functionality) to facilitate the disambiguation and linkage of Arabic entities with different languages. The entities that were considered include people, organizations, products, places, and events. Such entities have different names (lexical labels) in different languages, which is a major challenge faced when integrating knowledge from different sources, cultures, and languages. SIERA partners have defined a set of activities in task that aim **at resolving and linking Arabic entities with entities in multilingual portals** to leverage the knowledge across such multilingualism and culture diversities that the portals support.

In particular, resolving and linking Arabic entities to other Web entities across different languages through OKKAM ENS.

Organic.Edunet, in section 6, is a learning portal that provides access to thousands of quality multilingual digital learning resources on organic agriculture, agroecology and other green topics, such as ecology, sustainability, biodiversity, environment and energy. Organic.Lingua has been Arabized (its ontology, content, and searching and translation functionalities) to support Arabic search and retrieval.

Success Story: SMART Multilingual Tourist Guide in Bethlehem, in subsection 5.6. Based on the Arabization of the OKKAM platform, an innovative and pioneering electronic tourist guide in Bethlehem was developed and deployed in 10 heritage sites in the ancient city of Bethlehem. This was a result of successful and close cooperation of SIERA core and associate partners, as well as with the Palestinian Ministry of Tourism and Antiquities.
While walking in the old city, tourists can scan a QR code with smartphones and automatically listen to an audio guidance in their own language, as well as watch video and read further descriptions, about a certain heritage site.
This idea had big impact on the media: it was covered by 35 newspapers and radio stations, as well as through social media. The Palestinian Ministry of Tourism and Antiquities promised to expand this idea into more touristic and cultural places in Palestine.

Mapping Framework between WordNet and the Arabic Ontology has been established. This framework, presented in section 7, aims to support the alignment of concepts lexicalized in different natural languages, with a specific attention to aligning (mapping) Arabic concepts in the Arabic Ontology to their correspondent English concepts in WordNet. The framework provides the linguistic, theoretical and computational grounds for establishing such mappings. The framework was also experimented using different sources. A paper co-authored by several SIERA partners has been published at the Global Wordnet Conference (2014), where a panel discussion on this topic was also organized.

Project Information

Project Number	FP7–295006	**Acronym**	SIERA
Full Title	Integrating Sina Institute into the European Research Area		
Project URL	Sina.birzeit.edu/SIERA/		
Project Coordinator	Prof. Mustafa Jarrar mjarrar@birzeit.edu Birzeit Univerity, B.O.Box 14, Birzeit, Palestine		

Partner	Acronym/logo	Contact
Sina Institute, at Birzeit University, Palestine Page: http://sina.birzeit.edu/		• Prof. Mustafa Jarrar (Coordinator) • Prof. Adnan Yahya
Universidade Nova de Lisboa, Portugal Page: http://www.unl.pt/		• Prof. Christophe Roche • Prof. Rute Costa
Berlin-Brandenburg Academy of Sciences, Germany Page: http://www.bbaw.de/		• Prof. Christiane Felbaum
University of Trento, Italy Dept of Information Eng. and Computer Science Page: http://www.dit.unitn.it/		• Prof. Paolo Bouquet
University of Milano-Bicocca, Italy Page: http://www.unimib.it/		• Prof. Carlo Batini • Dr. Gianluigi Viscusi • Prof. Matteo Palmonari

Workpackage Structure

Person-Months per Participant

Participant number [10]		Participant short name [11]	Person-months per participant
1		BZU	37.00
2		FCSH-UNL	10.00
3		BBAW	8.00
4		UNITN	8.00
5		UNIMIB	5.00
	Total		68.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D2.1	Intermediate report on research setup	2	32.00	R	PP	18
D2.2	Report on Memorandums of understanding for PhD co-supervision.	2	2.00	R	CO	24
D2.3	Final report on research setup	2	33.00	R	PP	34
D2.4	Report on co-authored articles	2	1.00	R	CO	34
		Total	68.00			

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS4	A draft framework for matching WordNet and Arabic Ontology	2	12	
MS5	A draft of MICHAEL and KYOTO extension of Arabic concepts and sample content	2	12	
MS6	A draft of OKKAM extension with Arabic names and Entities	2	12	
MS7	Progress report on investigating the applicability of BZU Sina's APIs in MICHAEL and KYOTO	2	12	
MS8	Final framework for mapping WordNet and Arabic Ontology	2	30	
MS9	Final MICHAEL and KYOTO extension of Arabic concepts and sample content	2	30	
MS10	Final OKKAM extension with Arabic names and entities	2	30	
MS11	Final Report on investigating the applicability of BZU Sina's APIs in MICHAEL and KYOTO.	2	30	

Table of Content

1 Introduction

Task 2.2 "Set up joint research and cooperation" in WP2 aims to set up close research cooperation between the Sina institute and EU partners, in which exchanging knowledge and enhancing the cooperation capacity of BZU Sina will be facilitated. To achieve this goal concretely, 4 multilingual knowledge sharing portals (MICHAEL, KYOTO, OKKAM and Organic.Edunet) are selected as test bed to setup a sustainable research cooperating. The Task 2.2 is itself divided into 4 subtasks:

- Subtask 2.2.1: Investigating Arabic support in multilingual knowledge sharing portals
- Subtask 2.2.2:Extending multilingual thesauri/ontologies, for culture and ecology domains.
- Subtask 2.2.3: Framework for mapping between WordNet and Arabic Ontology
- Subtask 2.2.4: Resolving and linking entities and identities.

The deliverable is structured as the following:

Section 2 is dedicated to the presentation of the context and used resources. It includes a short presentation of: WordNet, the Arabic Ontology, and Arabic API's.

Section 3 is devoted to the Arabization of MICHAEL. Since the source code of MICHAEL was not available during the project, thus only our progress and current achievements are presented, such us the translation of MICHAEL's thesaurus, and its related theoretical backgrounds, standards, and tools.

Section 4 presents the mapping between KYOTO and the Arabic Ontology, which is the core step to the Arabization of SPARQL-based KYOTO portal. The full mappings are provided and discussed in this section, as well its importance, faced challenges, and open issues for the future cooperation are specified.

Section 5 presents the Arabization of OKKAM, including the Arabization of its ontology, interface, searching functionalities and the integration of BZU Arabic APIs, enriching its content with huge number of Arabic entities, and the resolving and linking of Arabic entities and identities. This section presents also a **success story on deploying a SMART multilingual tourist guide in Bethlehem.**

Section 6 focuses on the Arabization of the Organic.EduNet, including the Arabization its ontology, interface, searching functionalities, and enriching its content with well-annotated content from the agriculture domain.

Section 7 presents a revised and extended description of a mapping framework for ontologies lexicalized in different languages, with a specific attention to mapping between WordNet and the Arabic Ontology, and presents an experimental analysis for the framework.

 Section 8 presents the quality control and self-assessment procedure used for this deliverable.

As will be discussed and illustrated through out this deliverable, the presented efforts and achievements reflect a setup of close and sustainable research cooperation between EU and Palestinian scientists in the field multilingual knowledge sharing.

2   Description of Portals and of Used Resources

2.1 MICHAEL Portal

A multilingual Inventory of Cultural Heritage in Europe – is a European multilingual catalogue of digital cultural resources accessible online. The MICHAEL project was funded through the EU's eTen programme, to establish a new service for the European cultural heritage. The projects have established international online service, to allow users to search, browse and examine descriptions of resources held in institutions from across Europe[1]. The MICHAEL portal and Michael culture association home pages are shown in Figure 2.1 and 2.2, respectively.

"The Michael Culture Association is a not-for-profit organization, founded in July 2007. The association is established under Belgian law (AISBL - Association Internationale Sans But Lucratif). It was created following the Michael and Michael Plus European projects, which permitted to create the Michael European Portal, the Multilingual Inventory of Cultural Heritage in Europe. Its purpose was to continue the work on the portal beyond the period of the EU funding and to make the network of Michael partners sustainable.

Michael Culture Association has become a key actor in the promotion and valorization of the digital cultural content, and gathers a strong network of more than 100 institutions all over Europe." http://www.michael-culture.eu/presentation

Michael Culture leads the following actions:
- Animation of the network of exchange and cooperation
- Management of the network, development of new projects, creation of synergies with other networks
- A think-tank for digital cultural heritage field
- Strategic watch and prospective, organisation of workshops and working groups meetings
- Capitalisation and dissemination of the European projects of the Minerva network
- Promotion of best practices, communication, publications, guidelines…
- A services platform
- Innovative services for the general public, new services for the research

---

[1]For more information and references about MICHAEL, please check: http://www.michael-

Figure 2.1: MICHAEL portal home page
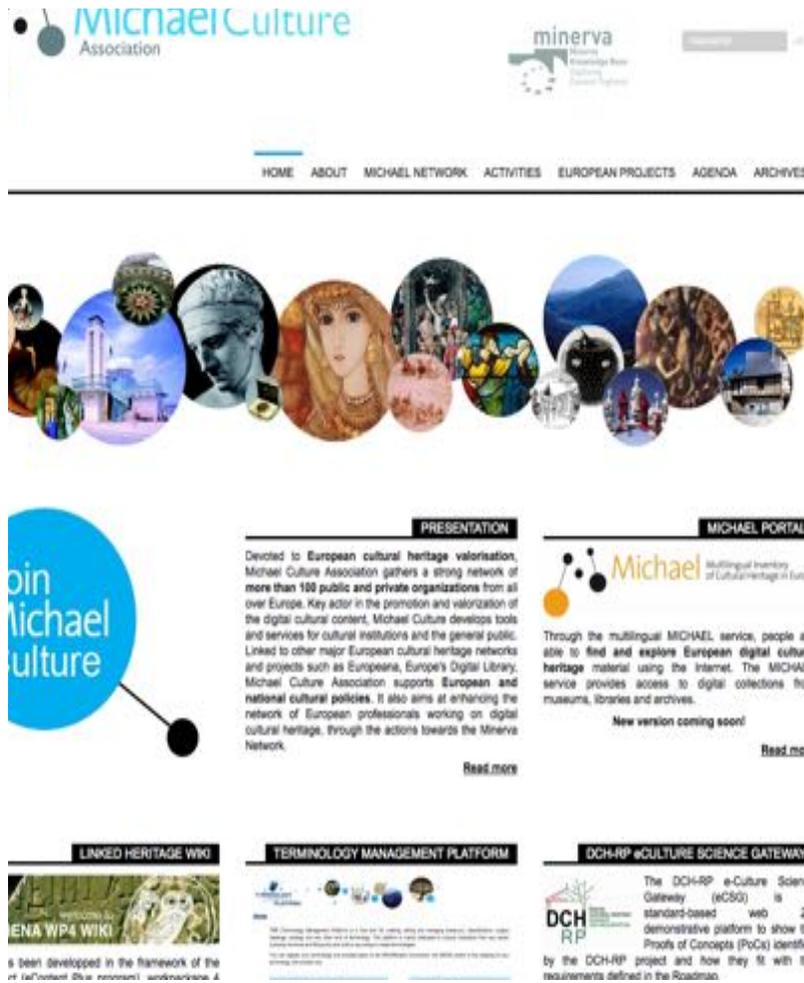
Figure 2.2: MICHAEL association home page

The Michael thesaurus is written in SKOS. SKOS – Simple Knowledge Organization System – is a W3C standard. As an interchange format, it provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. SKOS relies on concepts labeled with strings in one or more natural languages. It thereby enables a simple form of multilingual labeling (see Figure 2.3). SKOS will be presented in the part dedicated to the W3C Standards on Thesaurus and interchange format.
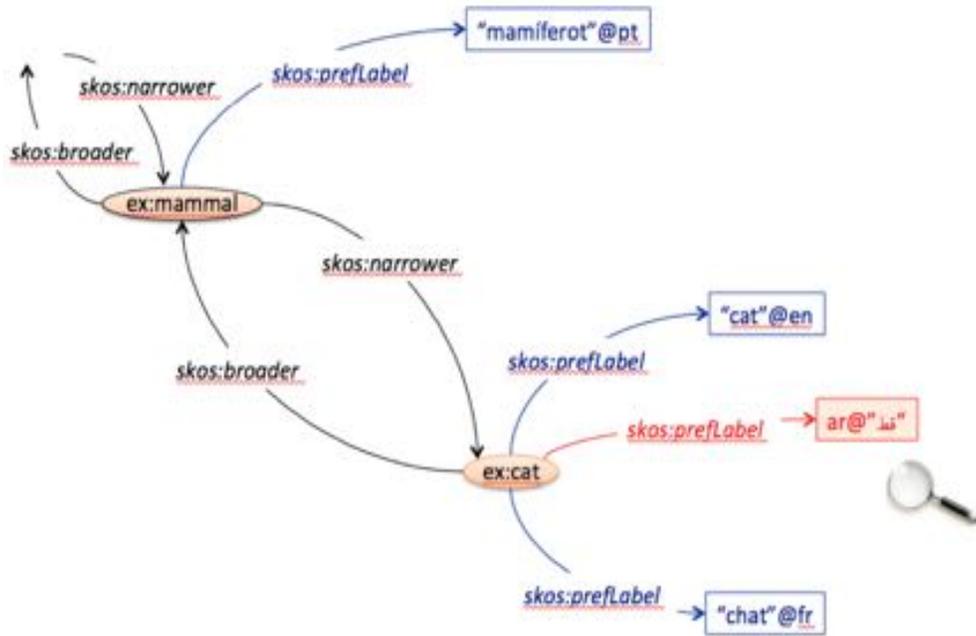
Figure 2.3: A multilingual SKOS representation. Arabic label "ar@قطة" for ex:cat shown in red

2.2 KYOTO

The KYOTO[2] project (ICT-211423) is a wiki-portal that provides a multilingual service to explore digital collections of environment and ecology objects and concepts, which also includes the resources and tools created for the KYOTO project. KYOTO developed an information and knowledge sharing system that relates text in various languages to a shared ontology in such a way that it enables the extraction of deep semantic relations and facts from text in a domain. The system establishes the communication and interpretation across languages and cultures and it supports building and maintaining the system by groups of people in a shared domain and area of interest. Figure 2.4 shows the KYOTO project home page.

Figure 2.4: KYOTO Project web page

KYOTO is an open platform that can be used to model and mine any kind of knowledge that is expressed in natural language text. A general representation model of text has been defined that can handle any textual structure. An ontological approach was considered to accommodate the extracted facts. The major role of the ontology in the KYOTO project is to provide a coherent, unified, stable frame of reference for different cultural and linguistic communities as well as different research communities.

KYOTO ontology covers the very general and abstract concepts in a clear, consistent manner that follows rigorous and explicit criteria. The KYOTO ontology (version 3)[3] has three distinct levels and is based on the DOLCE-Lite-Plus (DLP, a top level ontology). The DLP was adapted as the top-level; the middle level contains concepts that connect domain specific terms to concepts in the upper ontology. (e.g., Base Concepts, Units of measurement, and other qualities, etc). The third level is the domain specific level; this contains terms and concepts that are

---

[2] http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index.html

[3] http://weblab.iit.cnr.it/kyoto/xmlgroup.iit.cnr.it/kyoto/index8d0f.html?option=com_content&view=article&id=390&Itemid=155

pertinent to ecology. The three layers of the KYOTO ontology version 3 (final) are freely available for download in OWL format: KYOTO 3 Top level, KYOTO 3 middle level, and KYOTO 3 Domain level.

KYOTO ontology server stores and provides access to the KYOTO ontology. It is based on the VIRTUOSO server open source edition 4. Since the ontology is formalized in OWL-DL and serialized as a set of RDF triples, it is stored in a VIRTUOSO server as a RDF dataset and can be queried exploiting a SPARQL Endpoint.
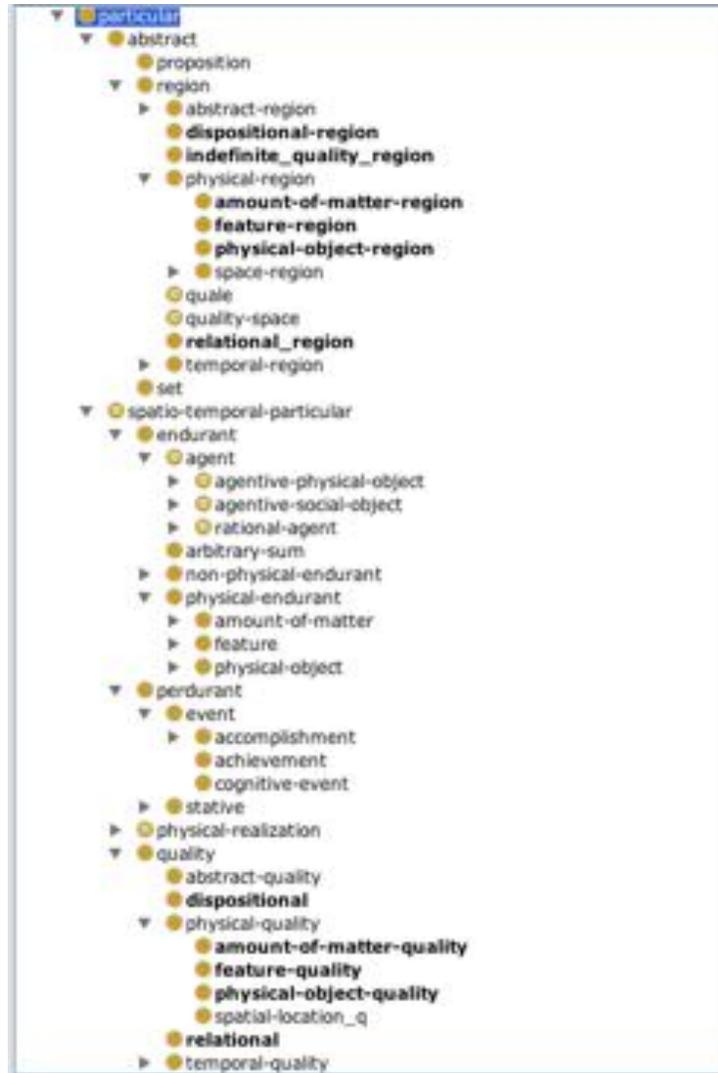

Figure 2.5: Screenshot of the KYOTO Ontology

2.3   OKKAM

OKKAM[4,5] (http://project.okkam.org/okkam-more) is spin-off of the University of Trento developing a collection of services, tools and methodologies for enabling an entity-centric management of data and knowledge in web-based applications and systems. Entity-centric or Web of Entities (as opposed to document-centric) means that the building blocks of the information space are entities (e.g., people, locations, organizations, events, products, etc.) and their relations, rather than digital documents and hypertext links.

OKKAM adopted a broad definition of notion "entity" which in principle covers all kinds of individual things. Accordingly, OKKAM has defined the 7 top-level categories: (Person, Organization, Event, Artifact type, Artifact instance, Location, Other) and to each class is attached a set of properties; for example, Person class has the properties (name, surname, age, gender, birth date, country, education).

The overall goal of OKKAM is to handle the process of assigning and managing unique identifiers for entities in the Web. These identifiers are global, with the purpose of consistently identifying a specific entity across system boundaries, regardless of who references this entity and from where. Figure 2.6 shows a high-level conceptual view of the OKKAM system and the interactions with its environment, where the end result is that all instances of the same entity (i.e., mentioned in different systems, ontologies, web pages, etc.) are assigned the same OKKAM identifier. Therefore, joining these documents and merging their information becomes a much more simple and effective process than before.

---

[4] P. Bouquet, H. Stoermer, and D. Giacomuzzi. *OKKAM: Enabling a Web of Entities*. In I3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.

[5] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008, number CSS-ICSC 2008-4-28-25. IEEE, August 2008.
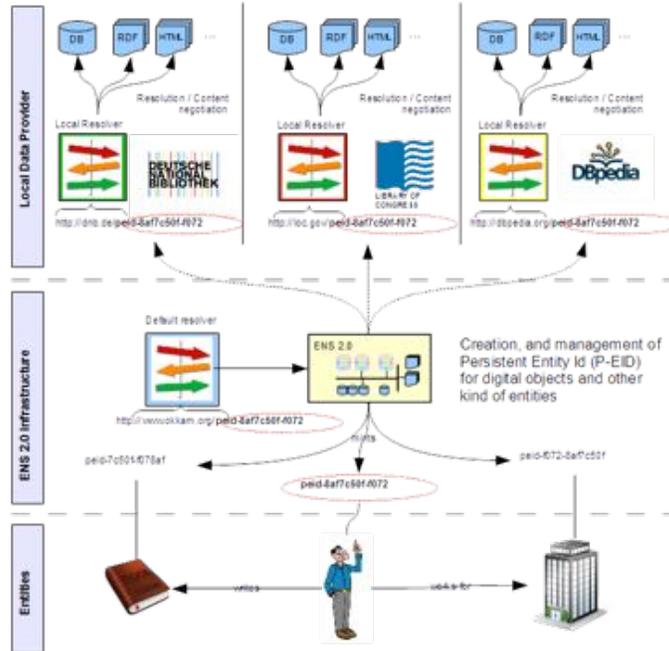
Figure 2.6: Schematic of OKKAM system and interactions[6]

OKKAM provides a scalable and sustainable infrastructure, called the Entity Name System (ENS). The ENS is a distributed service, which permanently stores identifiers for entities and provides a collection of core services (e.g., entity matching, ID mapping and resolution) needed to support their pervasive reuse. ENS enables Web applications to access the Web entities by getting and reusing OKKAM identifiers. Figure 2.7 shows the home page of OKKAM ENS.
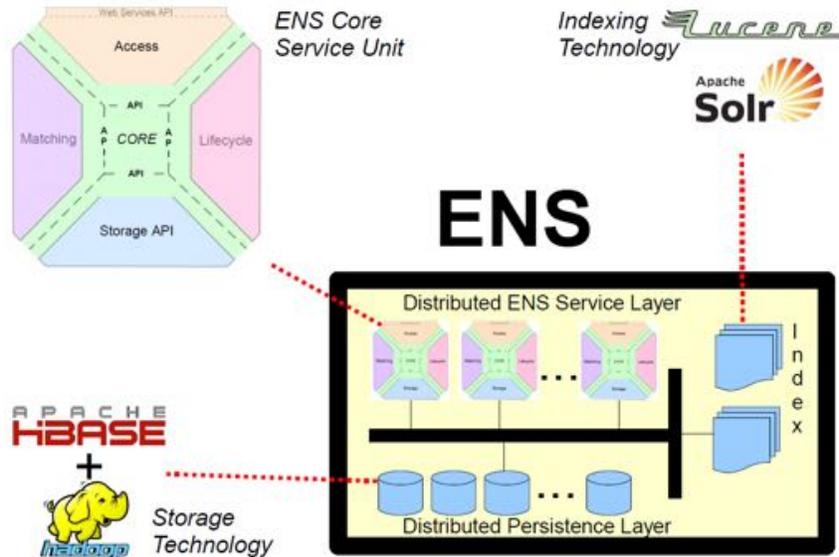
Figure 2.7:A view of the architecture of the Okkam Entity Name System

---

[6] http://project.okkam.org/deliverables/D6.1-EntityLifecycleManagement-Revised.pdf

ENS enables Web applications to access the Web entities by getting and reusing OKKAM identifiers. Figure 2.8 shows the home page of OKKAM ENS.
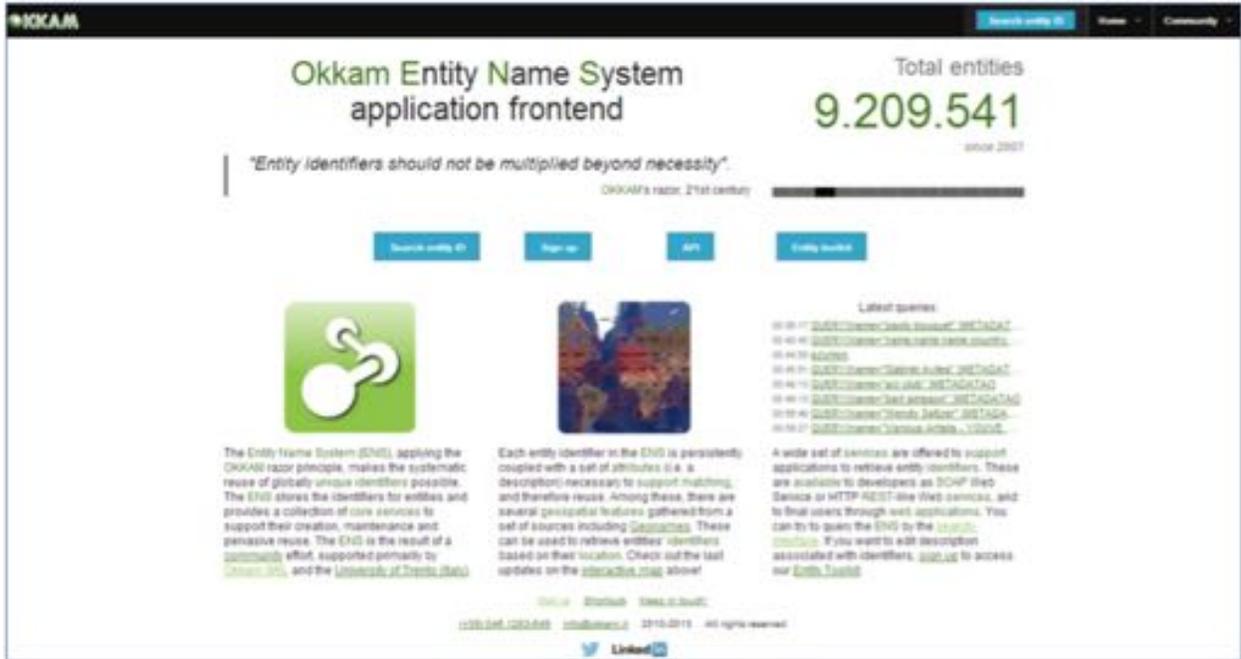
Figure 2.8: OKKAM ENS Web Interface

In order to support final users in trying the search services, OKKAM deployed a lookup interface available at http://api.okkam.org/search. The interface allows to select among the supported languages. As a first step, the user is expected to select the type of entity she is interesting into. Then, typing lookup attributes, attempting to compose a lookup request for an entity of interest. At first, the lookup interface suggests attribute names usefull for the lookup request (e.g. name, last_name, affiliation, etc.). These attributes are loaded from the Identification Ontology that will be presented more in detail in the following sections. Then the user types the values for those attributes types, composing a request. Then, the request is processed and sent to the Okkam ENS search APIs, and the result is returned to the user as shown Figure 2.9

Figure 2.9: OKKAM ENS Lookup Web Interface

2.4 Organic.Edunet

Organic.Edunet (www.organic-edunet.eu) is a learning portal that provides access to thousands of quality multilingual digital learning resources on organic agriculture, agroecology and other green topics, such as ecology, sustainability, biodiversity, environment and energy. Organic.Edunet facilitates the access, the usage and the exploitation of such educational content. The aim of Organic.Edunet is to provide learning resources which are appropriate for school and university level through a Web portal, as well as for vocational training and adult/lifelong learning thus targeting pupils, students, teachers, researchers, and general learners. Organic.Edunet achieved this by deploying a multilingual online federation of learning repositories, populated with multilingual quality content from various content producers.

The Organic.Edunet portal was re-engineered through the Organic.Lingua project7 to enhance its multilingual functionalities in order to facilitate its usage by users all over the world and cultures. Organic.Edunet portal user interface currently is available in nine languages and offers automatic metadata translation tools supporting eleven languages. The resources available through the portal are available in eleven languages (mainly English but also Estonian, French, German, Greek, Hungarian, Norwegian, Romanian, Russian, Spanish and Turkish); while the metadata for each resource are manually translated in up to eight languages and additional

---

7 http://www.organic-lingua.eu/

17

translations can be automatically provided by the portal's components. And recently Arabic was included in term of content and metadata. Figure 2.10 shows the home page of the Organic.Edunet portal. Users may search or browse for educational resources on Organic Agriculture and Agroecology using five different mechanisms: Text-Based Search, Browse, Semantic Navigation, Tag-Based Search and Search for educational scenarios. Although Organic.Edunet combines several technologies, including information retrieval and machine translation but its core functionalities are based a multilingual ontology, which has been also Arabized in SIERA.

Figure 2.10: the Organic.Edunet portal home page

2.5   WordNet

WordNet[8] is a manually constructed, large lexical database for English. It contains information about some 155,000 nouns, verbs, adjectives, and adverbs, organized in terms of their semantics. Specifically, words in WordNet that are similar in meaning are interlinked by means of pointers that stand for semantic relations. Formally, WordNet is a semantic network, an acyclic graph.

---

[8] Miller, G. A., and Fellbaum, C. (1991). Semantic Networks of English. *Cognition*, special issue, eds. Levin, B., and Pinker, S., 41.1-3: 197-229. Reprinted in: *Lexical and Conceptual Semantics*, eds. Levin, B. and Pinker, S., Cambridge, MA: Blackwell. 197-229.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM* 38: 39-41.

Fellbaum, C. (1998, Ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

WordNet superficially resembles a thesaurus but differs in several ways. First, WordNet interlinks not just word *forms*—strings of letters—but specific *senses* of words. Thus, words that are found in close proximity to one another in the network are semantically unambiguous. Second, WordNet explicitly labels the (numerous) semantic relations among words. Third, WordNet includes brief definitions and a short phrase illustrating the words' usage.

The main relation among words in WordNet is synonymy, as between the words *shut* and *close* or *car* and *automobile*. Synonyms are grouped into unordered sets, dubbed "synsets." Members of a synset are interchangeable in many, but not all contexts; the criterion for joint synset membership is merely that the words denote the same concept. Each of WordNet's 117,000 synsets is linked in turn to other synsets by means of a small number of "conceptual relations."

While WordNet represents synonymy straightforwardly in terms of synset membership, polysemy is reflected in the different synsets that a word form occurs in; membership in n synsets means that the word has n meanings, or is n-fold polysemous. In this way, each of WordNet's 225 000 form-meaning pairs is unique.

After the Princeton WordNet gained widespread popularity especially in the Natural Language Processing community, wordnets were built in a number of different languages. An important goal is to connect all wordnets to one another, so that equivalent words and meanings could easily be identified. Currently, some seventy wordnets have been or are being developed worldwide, including entire groups of languages such as EuroWordNet, BalkaNet, IndoWordNet and African WordNet.

The construction of crosslingual wordnets highlighted the need for a language-independent representation of concepts to which the lexicons (wordnets) of specific languages could be mapped. Using an actual language (like English) as the central hub is not the best solution, as it reflects the idiosyncracies inherent in each natural language; in particular, wordnets that are developed by translating the synsets may inherit such biases. Concepts and their relations to other concepts can be represented in formal structures, or ontologies, which are precise and unambiguous and lend themselves to reasoning and inferencing operations. SUMO (Niles and Pease 2003) is one such ontology, and it has been mapped to many wordnets, serving as its interlingua (Pease and Fellbaum 2009). Another ontology for wordnets of seven languages covering numbers of domain-specific terms has been developed in the context of the KYOTO project (http://www.kyoto-project.eu). Within the framework of this grant, the KYOTO and the Birzeit Arabic Ontology have been mapped (see Subtask 2.2.2.b "Extending the Kyoto's Ontology with Arabic Concepts and Objects").

2.6   Arabic Ontology

The Arabic Ontology[9,10,11] is an ongoing project, at Sina institute at Birzeit University, aims to build a lexical ontology for Arabic. The Arabic ontology is a formal representation of the concepts that the Arabic terms convey. For each term in the Arabic language, the set of its concepts (i.e., polysemy) are specified and given unique IDs, and formal semantic relations (such as subtype-of, part-of, and instance-of) are introduced between concepts. Synonymous terms that refer to same concept are grouped into Synsets. In this way the Arabic ontology can be seen as an Arabic Wordnet; however, unlike (the English) WordNet, the Arabic ontology is logically and philosophically well-founded, following strict ontological principles. For example, the Subsumption relation is a formal subset relation. The ontological correctness of a relation (e.g., whether "PeriodicTable SubtypeOf Table" is true in reality) in WordNet is based on whether native speakers accept such a claim. However, the ontological correctness of the Arabic ontology is based on what scientists accept; but if it can't be determined by science, then philosophers may accept it; and if philosophy doesn't have an answer then we refer to what linguistics accepts. Besides that, the OntoClean methodology is followed when dealing with instances, concepts, types, roles, and parts. Moreover, glosses[12] are formulated using strict ontological rules focusing on intrinsic properties. BZU-Sina Arabic Ontology Top Levels (AOTL)[13] provides a comprehensive overview of the meanings and formal and ontological properties of more than 250 Arabic core concepts. These core concepts of the Arabic Ontology (AO) capture the most abstract concepts in the Arabic language. They were formalized according to the kind of ontologically rigorous criteria that are used to develop top level ontologies such as BFO, DOLCE, and SUMO. Figure 2.11 presents the top levels of the Arabic Ontology, which is a classification of the most abstract concepts (i.e., meanings) of the Arabic terms. Only three levels are presented below for the sake of brevity. This is a novel approach to developing an electronic lexicon and promises much in the way of combining the best aspects of an electronic lexicon and a more rigorous ontology.

---

[9] http://sina.birzeit.edu/ArabicOntology/

[10] Mustafa Jarrar: Building A Formal Arabic Ontology. In proceedings of the Experts Meeting On Arabic Ontologies And Semantic Networks. Alecso, Arab League. Tunis, July 26-28, 2011.

[11] Mustafa Jarrar: Arabic Ontology, Lecture Notes. Sina Institute, Birzeit University, 2012.

[12] Mustafa Jarrar: Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In Proceedings of the 15th int. conference on World Wide Web, 2006.

[13] Mustafa Jarrar, Rana Rishmawi, Hiba Olwan: Top Levels of the Arabic Ontology. Technical Report. Sina Institute, Birzeit University, 2013.
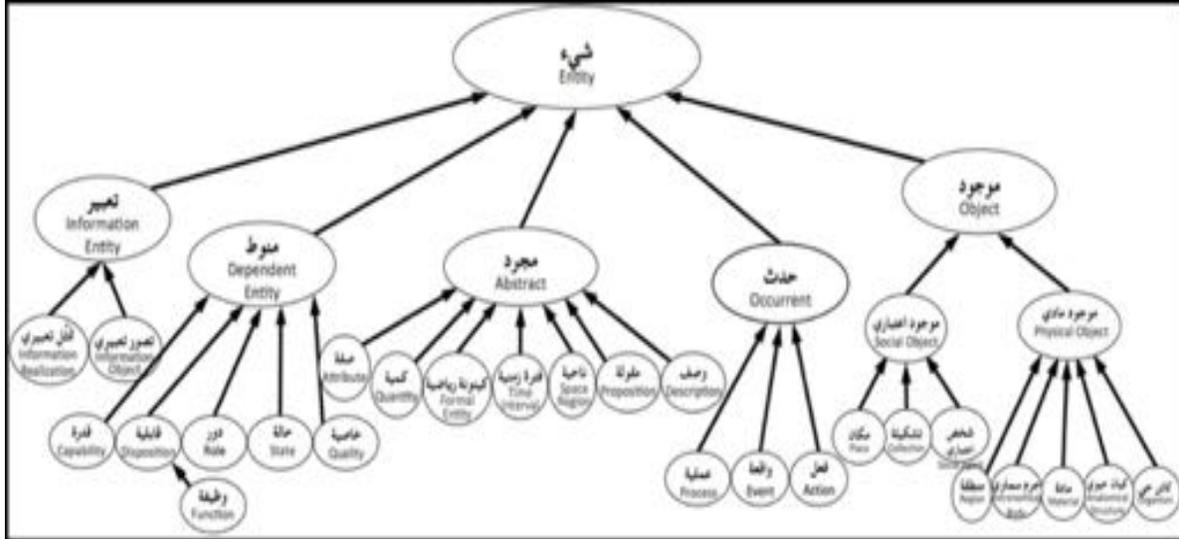
Figure 2.11: Part of the Arabic Ontology top levels

2.7 WOJOOD : Tools for Arabic Language Processing

BZU has developed several Arabic processing API's under the name of WOJOOD[14]. In WOJOOD project BZU has built several Arabic processing tools and databases for efficiently and easily retrieving and managing Arabic information on the Web. WOJOOD addresses the challenges Arabic poses for NLP and information retrieval, automatic Arabic document categorization, root extraction, language detection, and Arabic query correction, suggestion and expansion. It depends on statistical (corpus-based) approach based on contemporary data initially obtained from local newspapers, Arabic Wikipedia (for categorized databases), University and High School Students Names (for people names databases). The development of these tools was supported by BZU University through a research grant, by Google Inc. through a Google Research Award. Figure 12 shows the home page of WOJOOD.

Wojood tools were written in Java programming language designed and packaged to work as standalone services. That is, they can be easily integrated with any search engine built with Java; including the case of the proposed portals. Wojood provides a set of Arabic processing APIs we briefly describe them next, more details can be found at http://www.wojoodapis.com/.

---

[14] http://www.wojoodapis.com/

Figure 2.12: WOJOOD project home page

Tool 1 - Arabic language detector: detects whether the entered query is written in Arabic or not, that is it returns true if the entered query is Arabic and false if not, the tool is based on a large list of correct Arabic words with their frequencies

Tool 2 - Arabic Spell checking tool: Spell entered word, that it the tool receive a word as an input and returns a correction if it was misspelled, 4 possible corrections at max. The tool use a weighted equation to decide the correct outputs, the equation uses location (keyboard), sounds, similarity, shape similarity, relative frequency and Levenshtein distance to decide correct replacements from the same large set of correct words used in Tool1.

Tool 3 – Arabic Query Expansion: Expand the entered word to include other expanded list of it. For example play is expanded to played, playing … etc. (of course we are talking about Arabic words and the example is just to give an idea). The tool depends on a large rooted (stemmed) set of Arabic pairs {word, stem}, the tool simply extract the stem of the input word(s) and select all other words sharing same root (stem) from the table.

Tool 4 – Arabic Light Stemming: A tool that normalizes the input text by removing some unnecessary prefixes and suffixes from Arabic input.

Moreover, BZU Sina provided the partners with a set of utility tools that should support and enable the Arabic searching functionalities;

The Arabic POS tagger Web Service

BZU Sina provided an Arabic POS tagger web service based on the usage of Stanford Arabic POS tagger. The service is hosted on nlpsina.appspot.com. The API accepts the Arabic words as a string input, and at the server side the system split the string into words then calls a Stanford POS tagging[15] functionality, which then returns the part of speech of each word as a JSON format to the client side. Figure 13 demonstrates the usage of the POS tagger Web service. The shortcuts in the output are explained here http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.
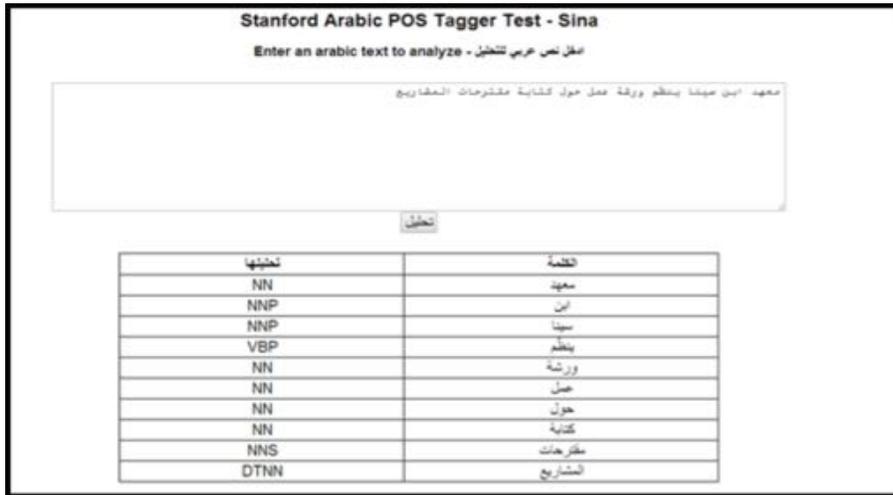

Figure 13: Arabic POS tagger Web Service

The Bilingual dictionary Web Service

BZU Sina also provided an access to its multilingual dictionary thorough a web services hosted at http://sina.birzeit.edu/mabuhelou/dic/. Sina dictionary were constructed by digitizing several lexical resources, about 200 thousand translation pairs are provided through the service. The service returns all the possible translations of the requested word either in English or Arabic as a JSON format to the client side. Figure 14 shows the (JSON) output of the translation of the Arabic word "سيارة" to the set of synonym English words {automobile, car, motor, motor car, sedan}, through the query

sina.birzeit.edu/mabuhelou/dic/query_json_back.php?lang=ar&word=سيارة&callback=jsonSina BiDicApi .
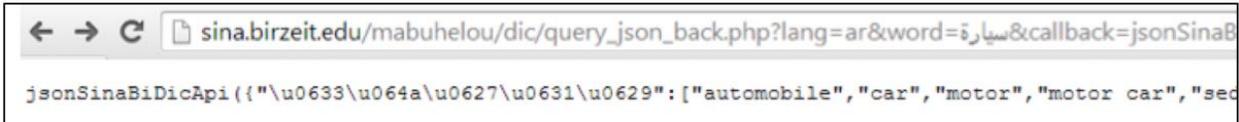
Figure 14: Arabic Bilingual Web Service out put for the query

---

3 Arabization of MICHAEL

This part is dedicated to the Arabization of MICHAEL. This original plans in SIERA's grant agreement was to extend MICHAEL portal with Arabic content, and searching functionality, which also requires extending MICHAEL's thesaurus. However, not all of these tasks were achieved due to the inaccessibility to the MICHAEL's source code. In what follows the progress we made towards the Arabization of MICHAEL.

3.1 Arabizing MICHAEL's Search

This subtask mainly concerns Arabic APIs and how to extend the Michael portal with Arabic search functionalities. The work which was done relies on the BZU's WOJOOD project whose goal is to build Arabic tools and databases for efficiently and easily retrieving and managing Arabic information. WOJOOD addresses the challenges Arabic poses for NLP and information retrieval, automatic Arabic document categorization, root extraction, language detection, and Arabic query correction, suggestion and expansion. It depends on statistical/Corpus-based approach based on contemporary data initially obtained from a local newspaper, Arabic Wikipedia (for categorized databases), University & High School Students Names (for people names databases). The results of this subtask were presented to a technical meeting in Paris on September 21st, 2012.

Due to the fact that the MICHAEL web portal was under development during the SIERA project, MICHAEL was not able to integrate the Arabic APIs in its portal. The work done in collaboration with Michael focused first on Michael's thesaurus (skosification, localization in Arabic and Portuguese) and second in setting up a national instance of the Michael portal dedicated to Palestine in order to integrate Arabic content.

3.2 Extending Michael's Thesaurus with Arabic Concepts and Content

This part is dedicated to **Thesaurus-oriented Multilingual Knowledge Sharing Systems**. It is illustrated with the Michael portal and Bethlehem content.

Objectives

This subtask aims at several objectives (the two first ones set down the main issues and describe the chosen approach when the last ones presents the context and the works carried out):

- First of all to identify the main issues raised by the thesaurus-oriented approach of multilingual knowledge sharing compared to the ontology-oriented approach taking into account the last revisions of ISO standards on terminology and thesaurus. The mapping issue, a fundamental issue, is studied from two points of view, ontology and thesaurus. We

will see that the extensional approach used in ontology matching cannot be directly applied in thesaurus;

- To propose a new paradigm combining, when it is possible, the two approaches. Since concept is the cornerstone of these systems, the ontoterminology approach - a terminology whose conceptual system is a formal ontology - appeared as a suitable and useful solution. Ontoterminology relies on the distinction between the conceptual dimension and the linguistic dimension - 2 non isomorphic dimensions - and easily allows to take into account multilingualism;

- To set up collaborations with European Projects which share similar objectives. Thus, a collaboration was set up with the Linked Heritage European project (ICT-PSP Project n° 270905). This project, dedicated to cultural content management, relies on multilingual thesauri (the SIERA project was presented at a Plenary meeting of the Linked Heritage project in November 2012 in Lisbon and during the Seminar on Multilingualism and Terminology organized by the Linked Heritage project and the French Ministry of Culture in Paris on April 18, 2013). This collaboration continues on the same subject of multilingual thesauri in the framework of the AthenaPlus European project, a CIP best practice network started in March 2013;

- The fourth section is devoted to the Michael thesaurus and to its localization in Arabic and Portuguese. The Michael thesaurus is written in SKOS. This W3C interchange format makes easy introducing new languages by defining new labels for concepts;

- The fifth section presents 3 multilingual-thesaurus oriented tools. The first one, developed by UNL in the framework of the SIERA project, is a multilingual-ontoterminology browser based on SKOS. It is dedicated to Michael's thesaurus and takes into account the Arabic language. The other two environments, TMP (Terminology Management Platform) and TMP2 (Thesaurus Management Platform version 2), were developed in the framework of the Linked Heritage and AthenaPlus projects. The last one (TMP2) relies on the ontoterminology paradigm where SKOS in only considered as an interchange format for thesauri;

- The sixth and last section is devoted to integrating Arabic content into Michael portal, it means integrating new Arabic concepts and objects related to Arabic Culture in relation to the Bethlehem Museum. This requires first to build the Bethlehem's Thesaurus in Architecture since objects must be not only described by attributes but first of all indexed by a controlled vocabulary. The second stage was to create a Palestine instance of the Michael portal in order to BZU and to the Centre for Cultural Heritage Preservation of Bethlehem to download data.

Technical meetings

In order to follow the progress of the project, several technical meetings were organized, one per year:

- The UNL works and UNL in-house methods and tools, as well as the Michael portal, were presented during the SIERA Kick-off meeting in Birzeit University on November 2011;

- A first WP2 Meeting dedicated to Michael's Thesaurus and TMP (Terminology Management Platform) was organized by UNL and Michael at the French Ministry of Culture on September 21st, 2012 in Paris;

- The second WP2 Meeting was organised by UNL in Lisbon on May 16th, 2013 devoted to Thesaurus and Bethlehem's Thesaurus Building.

- A last and third WP2 Meeting was organised by UNL and Michael in Paris on March 5th, 2014 in order to make the point at month 29.

Structure of this section

1. Thesaurus-oriented Multilingual Knowledge Sharing
2. ISO and W3C Standards
3. The Ontoterminology approach
4. Indexing and Information Retrieval
5. Mapping
6. Related European Projects
7. Coherency checking
8. Arabic localization
9. Portuguese localization
10. Multilingual Terminology and Thesaurus Editors

3.2.1 Thesaurus-oriented Multilingual Knowledge Sharing

Whatever the approach, Artificial Intelligence (Knowledge Representation), Linguistics and Lexical Semantics, Terminology, Thesaurus, Epistemology, we have to face 3 kinds of issues about:

- the "Reality" and the objects (or individuals) which populate it;
- the way we apprehend the "Reality", it means its conceptualization;
- and the way we speak about it;

We can sum up with the famous semantic triangle:
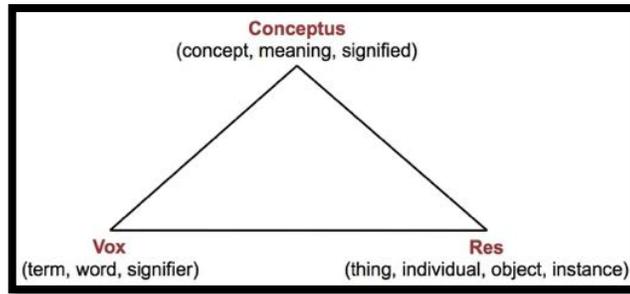
Figure 1: The semantic triangle

Nevertheless, the way each discipline tackles these issues can be very different. Thus, from a linguistic point of view the concept does not exist and can be reduced to a set of synonymous - linguistics is concerned with the relationship between words and their meanings -, when from an artificial point of view a concept, often called a class, is a shared structure of attributes whose values describe (more than define) the possible states of objects - Knowledge Representation is mainly concerned with the relationship between objects and concepts. From a logical point of view, a concept is a well formed formula (a unary predicate). When Terminology defines a concept as a unit of knowledge created by a unique combination of characteristics (essential[16] and delimiting characteristics), Thesaurus does not define concept in a formal language but put them into a network of concepts.

These different paradigms cannot be unified. This is not surprising since their goals are different. Linguistics is interested in Discourse when Artificial Intelligence is interested in Knowledge Representation. When the meaning of a word is built in discourse, the concept is a stable knowledge, and linguistic relationships cannot be reduced to linguistic translation of formal relationships between concepts. Thus, hypernym, a relation between words which can be multiple, is not a linguistic translation of subsumption, a relation between concepts which can be not multiple if it based on specific differences (Aristotelian definition). It is the reason why

---

[16] which requires a second order logic to be formalized, unlike to concepts in the Descriptive Logic.

the lexical network does not match with the conceptual network: "Saying is not Modelling"[17], making difficult the mapping of semantic networks and ontologies. At last, Knowledge Representation aims to represent objects when Thesaurus aims to index and classify them for information retrieval.

Nevertheless, the current revisions of the ISO Standards on Terminology and Thesaurus focus on the importance of concepts. It is the reason why ontology, from the Knowledge Engineering point of view, is one of the most promising perspectives for Terminology and Thesaurus. We will see that combining both ontology and terminology leads to the new paradigm of Ontoterminology.

3.2.2 ISO and W3C Standards

Today, it is no more possible to work on Terminology and Thesaurus ignoring the international Standards and the international works done on them. There are two main kinds of Standards useful for our purposes:

- The ISO Standards for Terminology and Thesaurus principles and methods.

ISO (International Organization for Standardization: http://www.iso.org) is an independent, non-governmental membership organization and the world's largest developer of voluntary International Standards. Their members are the national standards bodies of 163 member countries around the world. Its goal, since its creation in 1946, remains the same: 'to facilitate the international coordination and unification of industrial standards'. The ISO standards are developed by the people that need them, through a consensus process. Experts from all over the world develop the standards that are required by their sector. This means they reflect a wealth of international experience and knowledge.

- The W3C Standards for representation languages and interchange formats for publishing and linking data on the Web.

The World Wide Web Consortium (W3C: http://www.w3.org/) is an international community where member organizations and the public work together to develop Web standards. The Web of Data, one of the principles which guide the W3C's work, views the web as a huge repository of linked data. In this context, we are mainly concern with standard models for data interchange (RDF) and data organizing (SKOS, OWL) on which relies the architecture of the Semantic Web (figure below):

http://www.w3.org/standards/semanticweb/

---

[17] Roche, Christophe. 2007. "Saying is not Modelling." Natural Language Processing and Cognitive Science
[17](NLPCS) 2007, 47–56. ICEIS 2007, Funchal, Portugal, June 2007
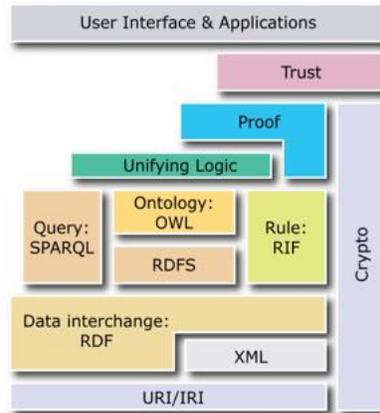
Figure 2: The Semantic Web Stack

The ISO Technical Committee TC 37 "Terminology and other language and content resources" is in charge of Standardization of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity. The ISO/TC37 is a horizontal committee whose goal is to provide guidelines for all other Technical Committees for their own terminological problems. The two main transversal standards on terminology principles and methods are the ISO 704 and the ISO 1087[18].

- ISO 704:2009. "Terminology work -- Principles and methods"

This standards establishes the basic principles and methods for preparing and compiling terminologies both inside and outside the framework of standardization, and describes the links between objects, concepts, and their terminological representations. It also establishes general principles governing the formation of designations and the formulation of definitions.

- ISO 1087-1:2000. "Terminology work -- Vocabulary -- Part 1: Theory and application

The main purpose of this international terminology standard is to provide a systemic description of the concepts in the field of terminology and to clarify the use of the terms in this field. This International Standard is addressed to not only standardizers and terminologists, but to anyone involved in terminology work, as well as to the users of terminologies. All the ISO Standards rely on the definitions of the ISO 1087-1. The definitions, from the terminology point of view, used in this Deliverable are the following (ISO 1087-1):

Terminology (1): set of designations belonging to one special language.
Terminology (2): science studying the structure, formation, development, usage and management of terminologies in various subject fields.
Term: verbal designation of a general concept in a specific subject field.

---

[18] Rute Costa is the ISO Convener of the ISO 704 and ISO 1087. Christophe Roche is the ISO Project Leader of the ISO 704 and ISO 1087.

Note: A term may contain symbols and can have variants, e.g. different forms of spelling.

Concept: unit of knowledge created by a unique combination of characteristics.

Note: Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background, which often leads to different categorizations. The definition of concept relies on the notion of essential and delimiting characteristics. From a logical point of view, the essential characteristic can be considered as a rigid predicate.

Characteristic: abstraction of a property of an object or of a set of objects.

Note: Characteristics are used for describing concepts.

Essential characteristic: characteristic which is indispensable to understanding a concept.

Delimiting characteristic: essential characteristic used for distinguishing a concept from related concepts.

Concept system: system of concepts - set of concepts structured according to the relations among them.

Hierarchical relation: relation between two concepts which may be either a generic relation or a partitive relation.

Generic relation: genus-species relation - relation between two concepts where the intension of one of the concepts includes that of the other concept and at least one additional delimiting characteristic.

Partitive relation: part-whole relation - relation between two concepts where one of the concepts constitutes the whole and the other concept a part of that whole.

Associative relation: pragmatic relation - relation between two concepts having a non-hierarchical thematic connection by virtue of experience.

Thesaurus

The ISO Technical Committee TC 46 "Information and documentation" is in charge of the Standardization of practices relating to libraries, documentation and information centres, publishing, archives, records management, museum documentation, indexing and abstracting services, and information science. The Standards on Thesaurus are split into 2 parts. The ISO 25964-1 and the ISO 25964-2.

- ISO 25964-1:2011. "Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval"

This part of ISO 25964 gives recommendations for the development and maintenance of thesauri intended for information retrieval applications. It is applicable to vocabularies used for retrieving information about all types of information resources, irrespective of the media used (text, sound, still or moving image, physical object or multimedia) including knowledge bases and portals, bibliographic databases, text, museum or multimedia collections, and the items within them. It is applicable to monolingual and multilingual thesauri.

ISO 25964-2:2013. "Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 2: Interoperability with other vocabularies"

> This part of ISO 25964 is applicable to thesauri and other types of vocabulary that are commonly used for information retrieval. It describes, compares and contrasts the elements and features of these vocabularies that are implicated when interoperability is needed. It gives recommendations for the establishment and maintenance of mappings between multiple thesauri, or between thesauri and other types of vocabularies.

Even if ISO standards rely (or should rely) on the ISO 1087-1, some of definitions used in Thesaurus are slightly different. The consequences of such differences can be important when we have to combine Terminology with Thesaurus. As a matter of fact, Thesaurus involves specific notions among which structured vocabulary, descriptor, preferred term, and facet analysis. These notions have to be defined.

Thesaurus: controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

Note: The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject. For this reason a thesaurus is optimized for human navigability and terminological coverage of a domain.

Structured vocabulary: organized set of terms, headings or codes representing concepts and their inter-relationships, which can be used to support information retrieval. Note: A structured vocabulary can also be used for other purposes. In the context of information retrieval, the vocabulary needs to be accompanied by rules for how to apply the terms. Various types of structured vocabulary will be addressed in ISO 25964-2, including classification schemes, subject heading schemes, etc.

Term: word or phrase used to label a concept.
Descriptor, Preferred term: term used to represent a concept when indexing.
Concept: unit of thought
Note: Unlike Terminology, Thesaurus does not aim to define concept as a unique combination of characteristics.
Facet: grouping of concepts of the same inherent category.
Facet analysis: analysis of subject areas into constituent concepts grouped into facets, and the subdivision of concepts into narrower concepts by specified characteristics of division.

W3C Standards
Standard model for data interchange on the Web relies on RDF (http://www.w3.org/RDF/), a Resource Description Framework for representing information in the Web. RDF is a graph-based data model relying on triples consisting of a subject, a predicate and an object (see figure below) meaning that a relation (the predicate) holds between the subject and the object.
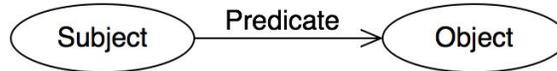
Figure 3: a RDF graph with 2 nodes (subject and Object) and a predicate linking them

SKOS (Simple Knowledge Organization System http://www.w3.org/TR/skos-reference/) provides a standard way to represent knowledge organization systems using the Resource Description Framework (RDF). SKOS provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. It relies on concepts labeled with strings in one or more natural languages. It thereby enables a simple form of multilingual labelling (see the figure below).
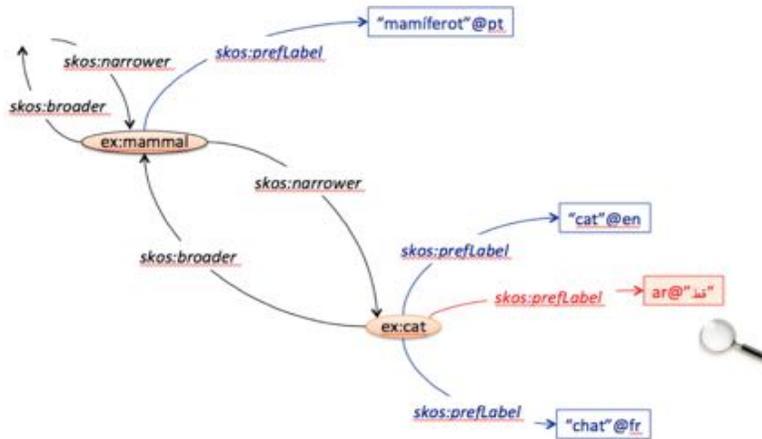
Figure 4: a SKOS graph

Nevertheless we have to bear in mind that SKOS is first of all an interchange format. SKOS is not a modelling language neither a formal knowledge representation language. Thus, the SKOS relationships cannot be guaranteed some logical properties. For example the skos:broader relationships and its inverse relationship skos:narrower do not distinguish between the "kind of" relationship and the "part of " relationship, the two fundamental relationships used in ontology building. Let us see the following example:

ex:country rdf:type skos:Concept;
skos:prefLabel "Country"@en.
ex:germany rdf:type skos:Concept;
skos:prefLabel "Germany"@en;
skos:broader ex:country.
ex:berlin rdf:type skos:Concept;
skos:prefLabel "Berlin"@en;
skos:broader ex:germany

Figure 5: Broader relationship

Of course Berlin is not a kind of country. It is the reason why the skos:broader and the skos:narrower are not defined as transitive in SKOS.

In a same way, the skos:broader and the skos:narrower are not defined as irreflexive in order to be able to import into SKOS ontologies written in OWL – the reflexive rdfs:subClassOf statement of OWL will be rewritten as a skos:broader relationship (the "subClassOf" relies on a set-inclusion meaning which is reflexive).

3.2.3 The Ontoterminology Approach

It is important to bear in mind that a terminology is not a thesaurus as well as a thesaurus is not a terminology. Indeed, the main goal of terminology is to define terms in relation to the domain conceptualization when the main goal of thesaurus is indexing content for information retrieval. Nevertheless both the latest versions of ISO standards on Terminology[19] and Thesaurus[20] emphasize the fact that concepts and terms must be separated as well as priority should be put on concepts since they are supposed to be linguistic independent - a concept is extralinguistic by definition. Thus, it should be possible to define concepts whose names are not descriptors putting forward the domain conceptualization. Here is an example where 'Germany' is an instance of 'Country' (Narrower Term Instantial) and 'Berlin' a part of 'Germany' (Narrower Term Partitive). 'Germany' and 'Berlin' are terms (descriptors) when 'Country' is not a term (descriptor). Putting forward concept directly leads to Ontology of the Knowledge Representation.

Country	Germany
NTI Germany	**NTP** Berlin

Ontology

Ontology, as defined in Knowledge Engineering, constitutes one of the most promising perspectives for Thesaurus and Terminology. According to the "famous" definition of Gruber[21], "An ontology is a specification of a conceptualization", more specifically "in the context of knowledge sharing. An ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents". In other words, an ontology is a system of concepts:

- linked by relationships: a kind of, part of, associative…

---

[19] Since the ISO/TC37 meeting in Madrid in 2012, the ISO 1087-1 and ISO 704 are under the process of revision in order to take into account new approaches coming from knowledge engineering and in particular ontology.

[20] the ISO 25964-1 published in 2011 and the ISO 25964-2 published in 2013 focus on concept and relationships between concepts.

[21] Gruber, Thomas R. 1992. "A Translation Approach to Portable Ontology Specifications." Knowledge Acquisition 5(2):199–220. DOI: 10.1006/knac.1993.1008

- defined and described by characteristics either essential or descriptive using.

Nevertheless, not all formal languages for concept definition are of equal merit[22]. They do not all offer the same functionalities nor the same guarantees. If the representation languages stemming from artificial intelligence have a long history and are human-readable, they do not always offer the required guarantees. For example, how can we accept the creation of a concept such as 'Metallic-Liquid-Element' in Mikrokosmos (subsumed by the 'Metal' and 'Liquid' concepts), which clearly causes confusion as to the nature (definition) and state (description) of something[23]?
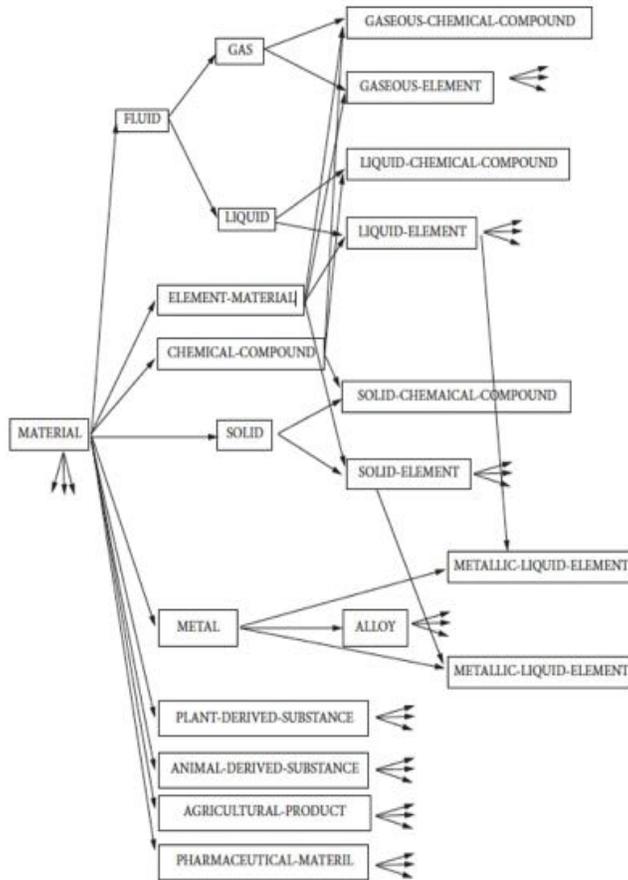
Figure 6: the Mikrokosmos Ontology

At the contrary, if formal languages, like description logic, allow to guarantee "good" properties in the logical sense they can raise epistemological problems. For example, the first order logic, which proposes only one paradigm (predicate) for knowledge representation, does not make it possible to distinguish between the different kinds of knowledge. For example it is not possible to distinguish, from the logical point of view, between HumanBeing (x), Mortal (x)

---

[22] C. Roche. 2014. "Ontological Definition" in "Handbook of Terminology", John Benjamins Publishing.

[23] Mercury is not a liquid metal, but a metal which is, under certain conditions of temperature and pressure, in a liquid state.

and Sick (x), even though those three unary predicates represent knowledge of a different nature – invariably true for the two first regardless of the possible worlds and contingent for the latter (Sick (x)). Taking into account essential characteristics like in Terminology requires higher order logic (in order to quantify characteristics) or modal logic (rigid predicates). Let us notice that some specific ontology environments like the Ontology Craft Workbench developed by the University of Savoie (see figure below) enables to define and manage essential characteristics according to the Aristotelian genus-differentia definition.
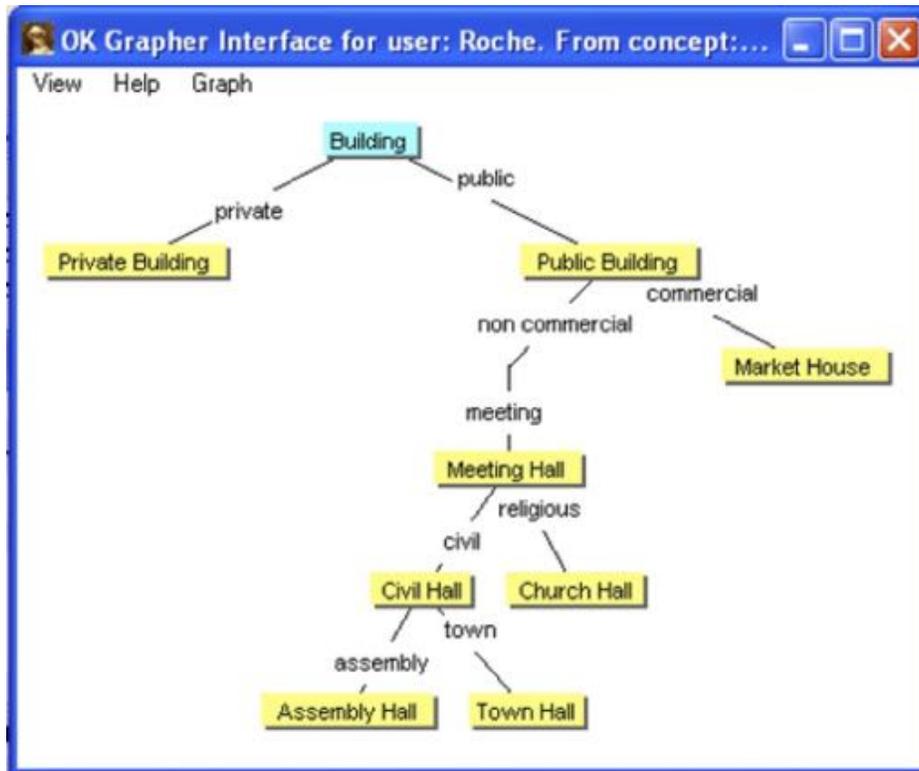
Figure 7: Ontology by "specific difference"

Ontoterminology

Combining ontology and terminology leads to the Ontoterminology[24] paradigm. This approach is based on: (1) a clear separation between the conceptual dimension – supposed to be common whatever the language - and the different linguistic dimensions – one per language; (2) an ontology-oriented approach for the conceptual model. The consistency of the terminology, and therefore its real usefulness and sustainability, is guaranteed by the ontology. The extra-linguistic representation of the conceptual system allows to define and link multilingual terminologies as illustrated by the following figure.

---

[24] Roche, Christophe. 2012. "Ontoterminology: How to unify terminology and ontology into a single paradigm". Eighth International Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turkey, May 21–27, 2012.
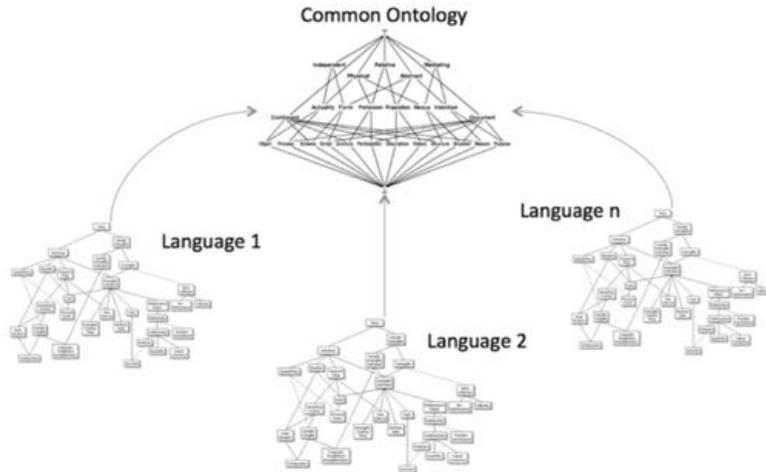
Figure 8: The ontoterminology approach

An ontoterminology is a terminology whose conceptual system is a formal ontology. The two non-isomorphic dimensions, conceptual and linguistic, are linked into a same paradigm (see figure below) leading to a Double Semantic Triangle putting into relations all the involved notions.
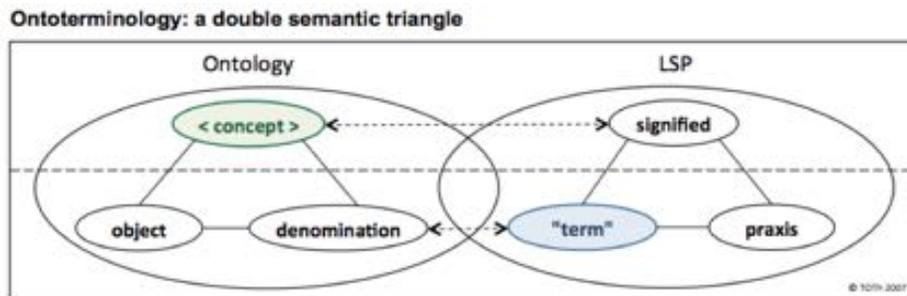

Figure 9: The double semantic triangle

By separating term and signified (a matter of natural language) on the one hand and the concept's name and its formal definition (a matter of the formal system) on the other, the double semiotic triangle does away with the constraint of bi-univocity). Only conceptualization, which is not a matter of linguistics but of science, is standardized. Linguistic diversity is preserved and, provided the conceptualisation of the world is shared, it becomes much easier to create multilingual terminologies.

3.2.4 Indexing and Information Retrieval

In separating the conceptual dimension from the linguistic one, ontoterminology defines a new approach for indexing and retrieval information. Terms, either preferred or non-preferred, are used to describe the content which will be classified onto the concepts denoted by the terms. Similarly, a document will be automatically analysed and classified onto the concepts denoted by the terms which appear in the documents. The figure below illustrates this approach.

Figure 10: Semantic and multilingual information retrieval principles

We can notice that there can be concepts without any descriptors attached to them - the ontology is useful for understanding the domain knowledge when descriptors are used for indexing documents. On the other hand we can look for information using only the ontology.

It is important to notice that a content can be indexed onto a concept in a given language and can get it back through another language. Furthermore, the search engine will be able to exploit the logical properties of the conceptual relationships in order to improve the search results.

The ASTECH FP6 project: "Advanced Sustainable Technologies for Heating and Cooling Applications" (2006-2009) used the ontoterminology-oriented approach for content management system. One of the goals of the project was to carry out a multilingual information retrieval system in renewable energy based on ontology and multilingual terminologies.

The figure below shows an example where the query is given in French through the ontology (concept <Transfert de chaleur>) and the first results are documents written in English indexed through English terms denoting the same extra-linguistic concept of <Transfert de chaleur>. It is an example of application where information retrieval is done through the ontology without taking into account the descriptors.


Figure 11: An application of ontoterminology-oriented information retrieval for content management system: the FP6 ASTECH project

3.2.5 Mapping

Mapping thesauri is a key step in extending thesaurus. Since the conceptual dimension of a thesaurus can be represented as an ontology, works done on ontology mapping can be useful for Thesaurus mapping. The structural models for mapping across thesauri are presented. The last paragraph presents the SIERA approach.

Ontology mapping

Mapping two ontologies consists in determining the relationships between their concepts.
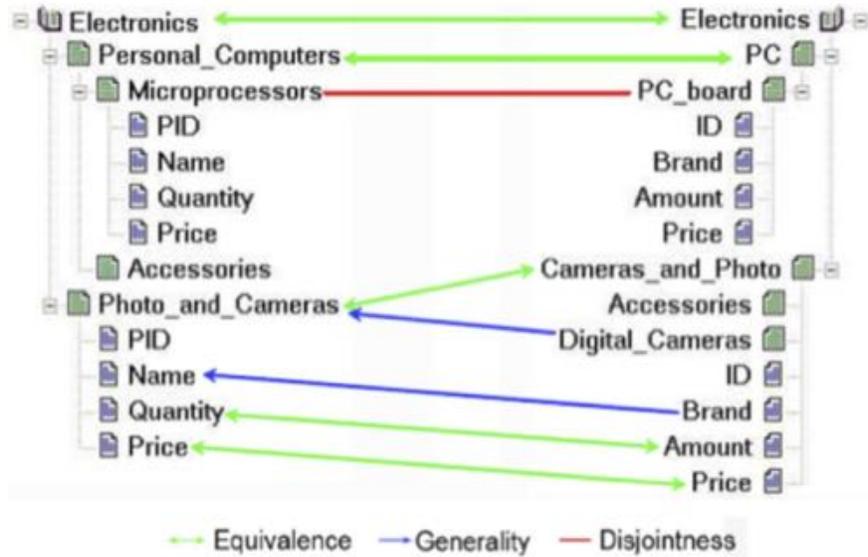
Figure 12: "Tutorial onSchema and Ontology Matching"
Pavel Shvaiko Jérôme Euzenat ESWC'05 – 29.05.2005

Two concepts can be either equivalent, more specific or more general, overlapping (sets of instances) or different. This information is embedded into a mapping element (c_1, c_2, R, n) where:

- c_1 and c_2 are concepts to be mapped;
- R is one of the relations $\equiv$ (equivalent), > (more general) , < (less general), $\cap$ (overlapping), $\perp$ (disjoint);
- n is a confidence measure.

The result of the matching process of 2 terminologies is the alignment of the terminologies defined as a set of mapping elements (see figure below).
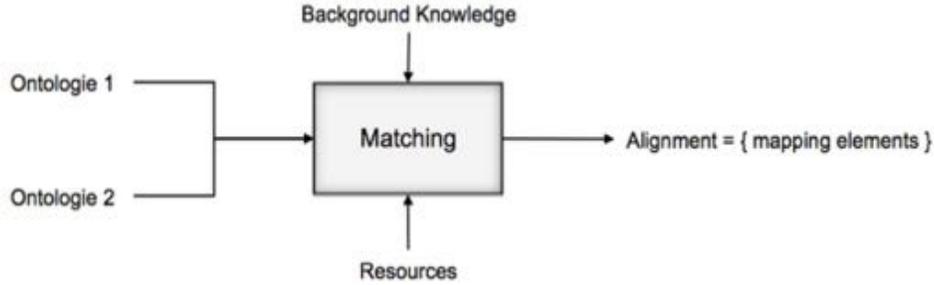
Figure 13: The matching process

Ontology matching has to face with three kinds of problems:
- *language mismatch* or language heterogeneity is when the ontologies are defined in different languages, e.g. frame language, logic, etc.
- *terminological (lexical) mismatch* is when the concept names are different including synonymy ("car", "automobile"), homonymy (same names with different meanings), terminological variation (same names with slight differences, ellipsis…)
- *conceptual mismatch* appears when there are different conceptualizations of the same domain, for examples "similar" concepts with different sets of instances (scope mismatch), "similar" concepts describing objects with different levels of details (granularity mismatch).

Mapping ontology can combine two types of methods, linguistic and semantic. The linguistic methods compare the concept names using string-based techniques based on the principle that the more the strings are similar, the more they denote the same concept. Thus, after lemmatization of the concept names, a distance between the two strings is calculated using different techniques: prefix (e.g. "int", "integer"), suffix (e.g. "phone", "telephone"), same letters or n-gram (common sequences of n characters), string metric (Levenshtein, Jaccard, etc.) Linguistic resources like dictionaries, thesauri or systems like WordNet can be mobilized to calculate the "distance" or the "similarity" between terms. For example "digital camera" is a hyponym of "camera". The semantic methods are split into two kinds of approaches, either extensional or intensional. The extensional approach relies on the sets of instances of the concepts postulating that the more two concepts have the same extension (common instances) the more they are identical[25]. The relations $\equiv$, $>$, $<$, $\cap$ and $\perp$ are defined according to the set operators. On the other hand, the intensional approach gathers methods based on the internal structure of concepts, i.e. on their attributes. These methods compare the attributes names (using linguistic methods) as well as their "semantics" i.e. the data type of the value of attributes (e.g. *date* and *working date*). The intensional approach also relies on the external structure of concepts, i.e. on relationships between concepts involving the graph structure (depth of concepts), the connected nodes (two nodes are all the more similar since their connected nodes are similar), etc. The following figure sums up the different methods.

---

[25] This mathematical definition of identity between sets (identity of extensions) is not really applicable in cultural heritage applications since a same set of objects can be "viewed", and then conceptualized, in different ways.
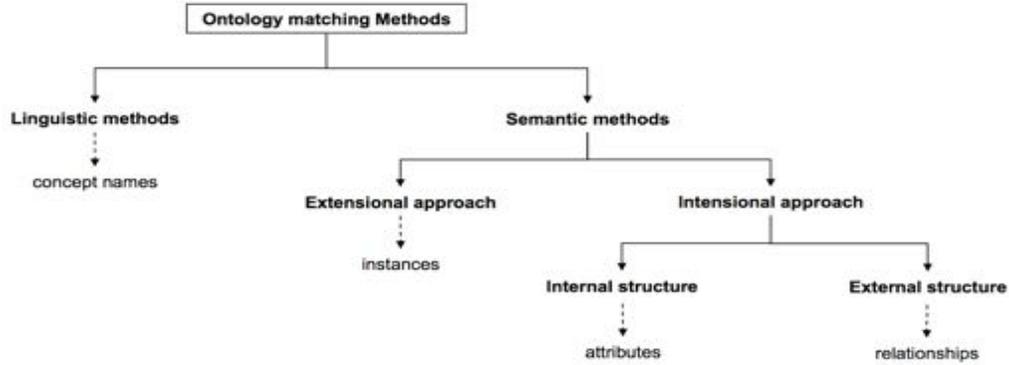
Figure 14: Ontology matching methods

Thesaurus mapping

Thesaurus mapping is the principal focus of the part 2 of the ISO 25964 (Part 2: Interoperability with other vocabularies). "the principal aim of interoperability between vocabularies[26] is to enable an expression formulated using one vocabulary to be converted to (or supplemented by) a corresponding expression in one or more other vocabularies" ISO 25964-2. In order to reach such a goal, mapping establishes relationships between vocabularies. We can identify several kinds of basic models for mapping thesauri.

The first one and the simplest, the Structural unity, is when all the thesauri share exactly the same structure of hierarchical and associative relationships between concepts. It is not really a mapping since terms coming from thesauri denote the same concepts. This model is directly managed by the ontoterminology approach as described by the figure 8.

The Direct-linked model (Figure 15) is when the thesauri do not share the same conceptualisation. It implies that direct mappings should be established between each thesauri.
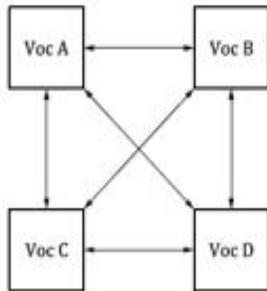
Figure 15: The Direct-linked model applied to 4 vocabularies
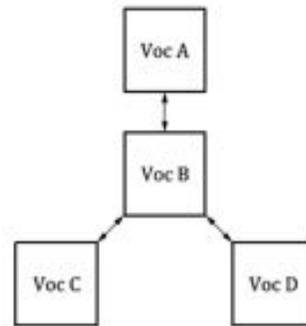
Figure 16: The Hub structure applied to 3 vocabularies mapped to a hub vocabulary

---

[26] Let us recall that a thesaurus can be defined as a controlled and structured vocabulary.

If there is a shared structure to which each thesaurus can be mapped, the Hub structure is a suitable architecture (Figure 16). Nevertheless, to achieve good quality mappings, the hub thesaurus needs to incorporate all the concepts present in all the thesauri.

The ISO 25964-2 gives some guidelines for choosing among these different structures.

Mapping establishes relationships between concepts. The three main types of mapping relationships to consider are equivalence, hierarchical and associative; respectively denoted by EQ (=EQ for Exact equivalence, for example "mad cow disease" =EQ "bovine spongiform encephalopathy"; ~EQ for Inexact simple equivalence, e.g. "potted plants" ~EQ "house plants"), BM and NM (BM for Broader Mapping, for example "streets" BM "roads"; NM for Narrower Mapping, e.g. "roads" NM "streets"), and RM (RM for Related Mapping, for example "e-learning" RM "distance education").

SIERA Thesaurus mapping

In thesaurus-oriented approach, a same item can be indexed by two different descriptors in two different thesauri (each of them describing a particular point of view). For example a same object will be classified under 'pottery' for a first thesaurus and under 'utensil' or 'handicraft' in another thesaurus. Thus, 2 concepts with the same extension can have different meanings and then cannot be mapped: 'meaning' and 'reference' have to be distinguished[27]. Furthermore, mapping thesauri is a means to find new sets of items, not more or less identical sets. It is the reason why we did not follow the extensional approach of ontology mapping, which relies on the fact that the more 2 concepts have common instances, the more they can be mapped.

The SIERA Thesaurus mapping relies first on a linguistic approach applied to the concept names based on the principle that the more the concept names are similar, the more the concepts are equivalent. The similarity between concepts corresponds to a distance between the concept labels. The distance is all the more smaller as the labels are similar. There are different measures of similarity between names (string of characters). One of the most popular string metric is the Levenshtein distance defined as the minimum number of necessary single-character edits to change one word into the other (insert, delete, substitute). For example, the distance between "examination" and "examination" is one since only one substitution is required ("e" for "a"). For longer strings (e.g. compound words), there are other measures like the Jaccard distance. We propose a semi-automatic mapping. It means that for each source concept (concept of the first terminology) the system proposes an ordered list of possible target concepts (concepts of the second terminology). The target concepts are ordered according to the string metric.

Distance

---

[27] "On sense and reference" G. Frege

The distance is based on a Levenshtein distance applied to all labels, either preferred labels or alternative labels, for all common languages. The minimum of all the calculated distances will be kept as the final distance. If there is no common language between the two terminologies, English will be chosen as pivot language (intermediary language). The labels of both terminologies will be translated into the pivot language.

Logical properties

The SIERA Thesaurus matching relies also on a semantic approach, in this case an intensional approach based on the hierarchical structure of the concept system. The logical properties of the hierarchical relationship of the two ontologies to be matched must be respected. It means that, for example, if a concept A exactly matches (=EQ) with a concept B, the narrower concepts of A can be aligned only with the narrower concepts of B.
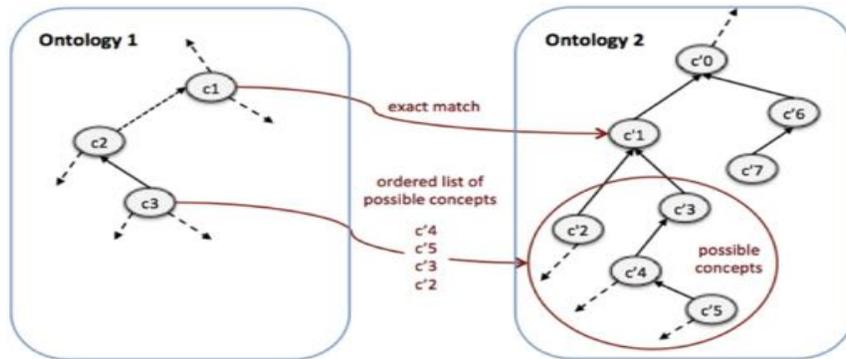
Figure 17: Thesaurus mapping

3.2.6 Related European Projects

Among the different European projects concerned with Thesaurus and Cultural Heritage, we have set up collaboration with 2 of them: the Linked Heritage Project and the AthenaPlus Project. The Michael association is one of the partners of these two projects.

Linked Heritage

A relationship was set up with the FP7 Linked Heritage Project (ICT-PSP Project n° 270905) and in particular with the WP3 about Terminology. One the goal was to study the interest of the conceptual approach in Thesaurus and of SKOS as interchange format for multilingual thesauri. One of the outcomes of this project, the TMP (Terminology Management Platform) has been used in SIERA. Web site: http://www.linkedheritage.org/

MICHAEL, the Associate Partner of SIERA for the Cultural Domain is also involved in the Linked Heritage Project. The SIERA project was presented at the 2012 Plenary meeting of the Linked Heritage project in Lisbon in November 2012. It was also presented (invited talk) during the "Seminar on Multilingualism and Terminology" in Paris on April 18th, 2013 organized in the framework of the Linked Heritage project.

Athena Plus

The collaboration initiated with Linked Heritage continues with the AthenaPlus Project, a CIP best practice network started in March 2013. The SIERA project is particularly interested in the new release of the TMP. The TMP 2 (Thesaurus Management Platform version 2) relies on the ontoterminology paradigm also chosen by SIERA and on the ISO Standards on Thesaurus. Web site: http://www.athenaplus.eu/

Michael Thesaurus

The first stage of the Arabic extension of the Michael Portal, or of any thesaurus-oriented portal, is its localization in Arabic i.e. to define an Arabic name for each concept of the thesaurus. The Michael Thesaurus is written in SKOS. It means that the processing of the thesaurus (checking, localization) depends on the SKOS properties.

3.2.7 Coherency checking

Before the localization, a verification phase was carried out in order to find any incoherence like missing concepts or more than one preferred label for one concept in a given language.

1. Missing concepts: Most of the missing concepts are due to spelling errors. For example:

-<skos:narrower
rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Buildings_and_the_build_environment"/>

and

-<rdf:Description
rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Buildings_and_the_built_environment">

2. More than 1 prefLabel for a same concept in a same language. SKOS does not verify the bi-univocity property "one concept - one preferred label". The checking phase allowed to identify several inconsistencies like: The concept: "Buildings_and_the_built_environment" has 2 different prefLabels:

- <skos:prefLabel xml:lang="en">Buildings and the built environment</skos:prefLabel>
- <skos:prefLabel xml:lang="en">Buildings and the build environment</skos:prefLabel>

3. A same prefLabel in a same language for 2 concepts. In a same way, the The checking phase allowed to identify different concepts with the same preferred Label. For example :

-<rdf:Description
rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Education_2">
        <skos:prefLabel xml:lang="en">Education</skos:prefLabel>
and:
-<rdf:Description
rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Education">
        <skos:prefLabel xml:lang="en">Education</skos:prefLabel>

3.2.8 Arabic localization

The Arabic localization of the Michael's thesaurus consists in defining new labels in Arabic for SKOS concepts. It relies on the hypothesis that for each Michael concept there is an Arabic label.

Figure 18: Arabic localization of the Michael Thesaurus

3.2.9 Portuguese localization

Based on the same principle, the Portuguese localization of the thesaurus is done by defining new Portuguese labels for SKOS concepts.

Figure 19: Portuguese localization of the Michael Thesaurus

3.2.10 Multilingual Terminology and Thesaurus Editors

We saw that the thesaurus-oriented approach for multilingual knowledge sharing requires a new paradigm combining both ontology and terminology. Ontoterminology is one of them. This paradigm allows to share a same conceptualization between different natural languages without any direct connection between terms in the different languages. This ontoterminology-oriented approach requires specific environments or at least environments which allow to distinguish between the conceptual dimension and the linguistic dimension. This section presents three software environments. The first one, developed by UNL, is ontoterminology-oriented when the second one, developed during the Linked Heritage project, is dedicated to Thesaurus written in SKOS. The last one, currently developed in the AthenaPlus project, is based on the ontoterminology paradigm taking into account the last version of the ISO 25964-1 and ISO 25964-2 Standards.

Multilingual Ontoterminology Browser

During the SIERA project, UNL developed "OTe for SKOS", a software environment dedicated to the Michael Thesaurus taking into account: - Multilingualism and Arabic language; - Ontoterminology paradigm; - SKOS format (since Michael's Thesaurus is written in SKOS); Browsing functionalities (dynamic navigation) based on the conceptual system; - Management of the linguistic and conceptual dimensions. The "OTe for SKOS" environment developed by UNL provides different ways of navigation, mainly a tree view based on the generic hierarchical relationship between concepts (broader/narrower relationship), but also an interactive concept map allowing navigation between concepts through the conceptual relations.

Figure 20: The tree view of concepts in Arabic   Figure 21: The interactive concept map in Arabic
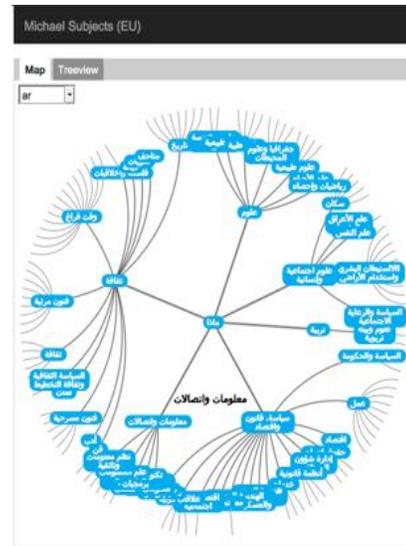
"OTe for SKOS" allows to access to the 2 dimensions, conceptual and linguistic, of an ontology-oriented thesaurus whatever the language, including Arabic language. The "linguistic card" displays all the linguistic information attached to the concept in a given language, e.g. the term (preferred label) as well as alternative terms like synonymous, acronyms, linguistic variations… (alternative labels), definition in natural language, comment and notes. The two following figures present the terms in English an in French denoting the <Social and human sciences> concept.

Figure 22: The linguistic card in English associated to the <Social and human sciences> concept

Figure 23: The linguistic card in French associated to the <Social and human sciences> concept

The "conceptual card" displays the concepts linked to the selected concept, i.e. the <Top concept>, for example the domain the selected concept belongs to, the <broader concepts>, the <narrower concepts> as well as the <related concepts> (linked by the related-to relationship).

Figure 24: The <Social and human sciences> concept card in English

All these functionalities are integrated into "OTe for SKOS" environment. The following figure is a hard copy of the software. The left pane allows one to browse inside Michael's Thesaurus through the ontology. The right pane displays the linguistic or the conceptual card associated to the selected concept in the left pane.
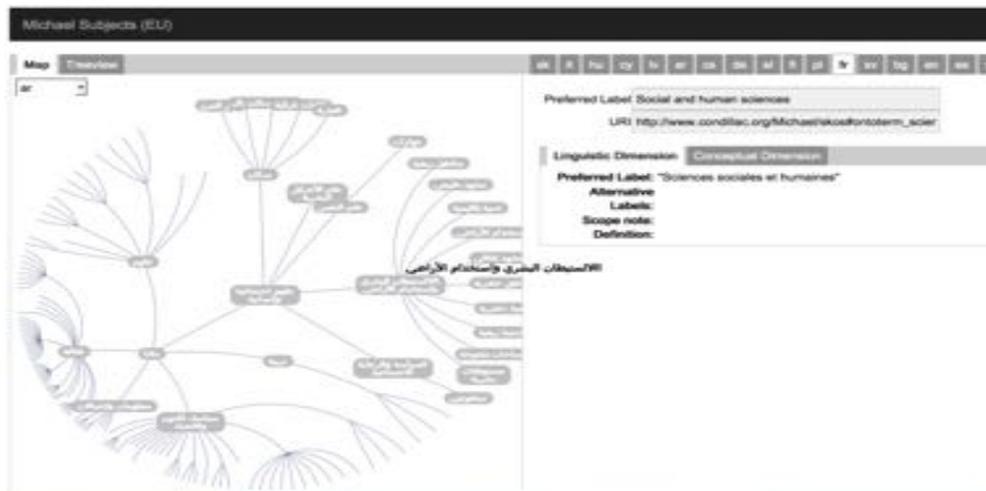
Figure 25: The <Social and human sciences> concept card in English

It is important to notice that the language used for the conceptual browser and the language used for the conceptual and linguistic cards are independent (in the previous figure, Arabic for the ontology map and French for the linguistic card).

Terminology Management Platform

During the SIERA project, a collaboration was set up with the Linked Heritage Project on multilingual thesauri and multilingual terminologies. The TMP, for Terminology Management Platform, was developed during the Linked Heritage project. The TMP is SKOS-oriented. It means that SKOS is not only used as an interchange format (import and export), but also as the internal representation of terminology and thesaurus. It also the reason that the TMP interfaces

are based on SKOS. The TMP was presented at two WP2 Technical meetings in Paris in 2012 and in Lisbon in 2013. A TMP account was created for SIERA (username and password) for building the Bethlehem thesaurus. http://www.culture-terminology.org (user: siera, password: siera)

Figure 26: The TMP platform illustrated with the Michael's Thesaurus

Thesaurus Management Platform version 2

Based on the returns on experiences on the Linked Heritage project and on the SIERA project, a new version of the TMP (TMP2) is currently under development in the framework of the AthenaPlus project. SKOS is no more the internal format, but only one of the available interchange formats. The TMP2 core is the Ontoterminology (OTe: OntoTerminology engine) and the functionalities those of the ISO 25964 on Thesaurus. Unlike the first version of the TMP, this new release takes into account instances (proper names), the part-of relationship as well as micro thesauri and facets (as they are defined in ISO 25964-1).

Figure 27: The Thesaurus Management Platform release 2

3.2.11 Integrating Arabic Content into Michael Portal

Integrating new content (the experimental sample of 1000 objects related to Arabic culture and ecology collected by BZU Sina) of a new domain (Architecture in Bethlehem) in a thesaurus-oriented content management systems requires:

- Building the Thesaurus of the domain, in order to define the domain ontology and the descriptors (terms) associated to the concepts;
- Indexing the objects using the descriptors (terms) of the thesaurus;
- Integrating (or to maping with) the Bethlehem's Thesaurus into the Michael's Thesaurus;
- Uploading the objects into the Michael Portal.

This section starts with the description of the objects related to Arabic culture (description of the Bethlehem's buildings of the historic town). Then the stages of "thesaurization" of Bethlehem data is presented. The section ends with the presentation of the idea of a Palestinian instance of the Michael portal for integration of data.

Objects related to Arabic Culture

The experimental sample of objects related to Arabic culture collected by BZU Sina in collaboration with the Centre for Cultural Heritage Preservation of Bethlehem is a set of building descriptions of the historic town of Bethlehem (867 buildings).

Each building is described by a set of attributes with values like the parcel number or the type of property (figure 28.a and figure 28.b). These attributes are organized into different types: General Information, Type of Property, Composition, Period of Construction, Building Description, etc. (see figures below).

General Information

Building No. **89**	Sheet No. **55b**
Parcel No. **93**	Block No. **28023**
Number of Inhabitants: **none**	Building Name: **Dar Al Ghazzawi (Issa Dauod Al Ghazzawi)**
Classification: **B/2**	

Figure 28.a: Bethlehem Data Description

About the Building

Type of Property: ☑ Private Ownership ☐ Public Ownership (governmental) ☐ Religious Institutions	Period of Construction ☐ Before Ottoman Period (Church of the Nativity and its Surrounding) ☐ Ottoman Period (1517-1917) ☑ British Mandate (1917-1948) ☐ Jordanian Jurisdiction (1948-1967) ☐ Israeli Occupation Period (1967-1995) ☐ Palestinian Authority (1995-today) ☐ Other	
Year of Construction: **1924**	Connection to Sewage network: ☑ Yes ☐ No	
Availability of water well: ☑ Yes ☐ No	If yes, indicate capacity (m³): **35**	
Indicate if used: ☐ Yes ☑No	Comments: **abundant building**	
Urgent need for intervention (residential)	☐ Yes	☑No

Figure 28.b: Bethlehem Data Description

Building Composition

Type of Construction	Major Additions
☐ Minor structure (Tin, Corrugated Sheets, roof tiles, etc…) ☐ Vaulted Ceiling (Barrel of Cross) ☐ I-Beam ☑ Composition of Both (i-beam and vaulted) ☐ Concrete Structure ☐ Timber and Roof Tiles (Churches and Convents)	☐ 21% - 50% ☐ 51% - 75% ☐ 76% - 100% ☐ 101% - 150% ☐ More than 151%

Figure 28.c: Bethlehem Data Description

Building Development

Composition/Morphology of the Building

☐ Minor structure (e.g. tin, corrugated sheets, roof tiles, etc…)

☐ Traditional building and/or agricultural cell

☐ Traditional residential complex (*Hosh*)

☐ Traditional building derived from synchronic project referred to local culture/architectural vocabulary

☑ Traditional building derived from synchronic project referred to European culture/architectural vocabulary

☐ Modern building derived from extensive restoration of traditional building

☐ Modern building coherent with pre-existing urban plot (material/floor number/building scale/shape/composition)

☐ Imposed and minor additions

☐ Specialized building/complex

☐ Commercial building

Figure 28.d: Bethlehem Data Description

All those information about the buildings are available as a RDF file (figure 29) where each attribute is described by a specific tag, for example:

<bb:PARCEL_NUM>40</bb:PARCEL_NUM>

```
- <rdf:Description rdf:about="building#8">
    <bb:BUILDING_N>8</bb:BUILDING_N>
    <bb:FID_>0</bb:FID_>
    <bb:SHEET_NUMB>45a</bb:SHEET_NUMB>
    <bb:PARCEL_NUM>40</bb:PARCEL_NUM>
    <bb:BLOCK_NUMB>28014</bb:BLOCK_NUMB>
    <bb:NUMBER_OF>0</bb:NUMBER_OF>
    <bb:BUILDING_1>1</bb:BUILDING_1>
    <bb:TYPE_OF_PR>2</bb:TYPE_OF_PR>
    <bb:PERIOD_OF>1889</bb:PERIOD_OF>
    <bb:YEAR_OF_CO>1</bb:YEAR_OF_CO>
    <bb:CONNECTION>1</bb:CONNECTION>
    <bb:AVAILABILI>24m3</bb:AVAILABILI>
    <bb:WELL_CAPAC>0</bb:WELL_CAPAC>
    <bb:URGENT_NEE>4</bb:URGENT_NEE>
    <bb:BUILDING_C>2</bb:BUILDING_C>
    <bb:TYPE_OF_CO>5</bb:TYPE_OF_CO>
    <bb:ADDITIONS>4</bb:ADDITIONS>
    <bb:MORPHOLOGY>0</bb:MORPHOLOGY>
    <bb:A_AESTHETI>0</bb:A_AESTHETI>
    <bb:A_HISTORIC>0</bb:A_HISTORIC>
    <bb:A_SOCIAL_V>0</bb:A_SOCIAL_V>
    <bb:A_AUTHENTI>0</bb:A_AUTHENTI>
    <bb:A_LOCATION>3</bb:A_LOCATION>
    <bb:C_EXTERNAL>4</bb:C_EXTERNAL>
    <bb:C_INTERNAL>4</bb:C_INTERNAL>
    <bb:C_STRUCTUR>4</bb:C_STRUCTUR>
    <bb:C_OVERALL>3</bb:C_OVERALL>
    <bb:C_AIR>3</bb:C_AIR>
    <bb:C_LIGHT>3</bb:C_LIGHT>
    <bb:C__ENVIRON>88</bb:C__ENVIRON>
    <bb:SHAPE_LENG>269</bb:SHAPE_LENG>
    <bb:SHAPE_AREA>708906.34176</bb:SHAPE_AREA>
    <bb:X_COORD>3510289.35611</bb:X_COORD>
    <bb:Y_COORD>31.7089554235574</bb:Y_COORD>
    <bb:Lat>35.204513004193</bb:Lat>
  </rdf:Description>
```

Figure 29: Bethlehem Data Description in RDF

Bethlehem Thesaurus and Data

Each Bethlehem's building description is defined as a set of attributes with values (more than 867 buildings, each of them described by 36 attributes). It is a more artificial intelligence-oriented description than thesaurus-oriented.

```
535    <rdf:Description rdf:about="building#134">
536        <bb:BUILDING_N>134</bb:BUILDING_N>
537        <bb:FID_>0</bb:FID_>
538        <bb:SHEET_NUMB>65a</bb:SHEET_NUMB>
539        <bb:PARCEL_NUM>144</bb:PARCEL_NUM>
540        <bb:BLOCK_NUMB>28024</bb:BLOCK_NUMB>
```

Figure 30: Attributes of Bethlehem Data Description

In order to take into account these data in a thesaurus-oriented content management system, they must be indexed by a controlled and structured vocabulary (the Bethlehem Thesaurus).

The solution consists in completing the RDF building descriptions with the Bethlehem thesaurus terms using the metadata of Dublin Core, for example the 'subject' metadata. In the example below, the building #134 is indexed by the "Public Building" descriptor.

```
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dcam="http://purl.org/dc/dcam/">
<rdf:Description rdf:about="building#134">
        <dcterms:subject>
                <rdf:Description>
                        <dcam:memberOf rdf:resource="http://Bethlehem/Thesaurus/Building"/>
                        <rdf:value>Public Building</rdf:value>
                </rdf:Description>
        </dcterms:subject>
         <bb:BUILDING_N>134</bb:BUILDING_N>
        <bb:FID_>0</bb:FID_>0
        ....
```

Bethlehem Thesaurus

The Bethlehem Thesaurus is about the historic town of Bethlehem. It is a kind of Architecture Thesaurus. Several Architecture thesauri were studied among which the English Heritage Monument Type Thesaurus and EuroVoc, the Multilingual Thesaurus of the European Union.

EH Monument Type Thesaurus

- http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes_no=1

Figure 31: English Heritage Thesaurus

Figure 32: EH Monument Type Thesaurus

EH Monument Type Thesaurus

http://eurovoc.europa.eu/drupal/?q=request&view=mt&mturi=http://eurovoc.europa.eu/100273
&language=en

Figure 33: The EuroVoc Thesaurus

The Building Ontology of Bethlehem

Following the Ontoterminology approach of Thesaurus, the first stage is to define the domain ontology using an ontology editor (let us recall that SKOS is not a modelling language).

The figures below represents one of the views of this ontology defined with 2 different concept system editors. The first one is based on relation theory, the second one on descriptive logic (Protégé). The 'Type of Property' attribute and its values 'Private Ownership', 'Public Ownership' and 'Religious Institution' are represented as concepts linked by the subclass (is-a-kind-of) relationship.
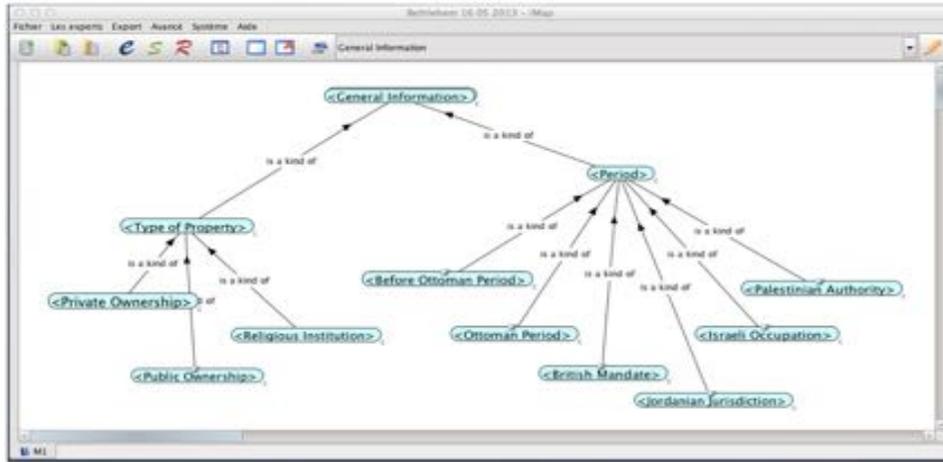
Figure 34: iMap Ontology editor
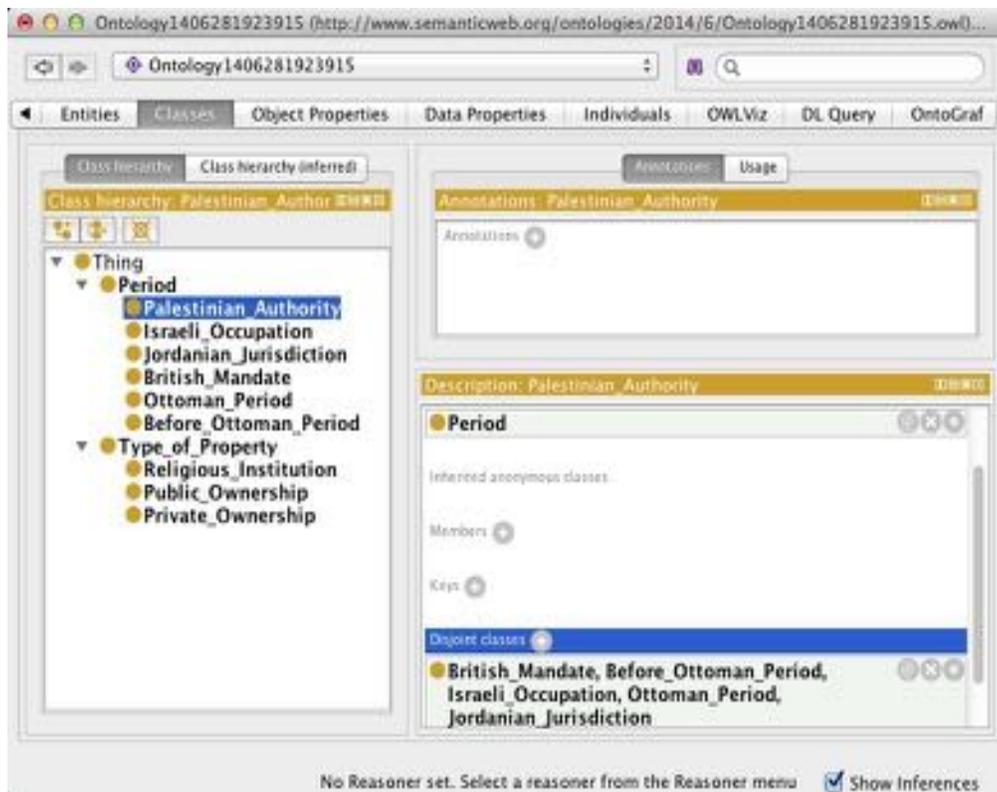
Figure 35: The Protégé Editor

Building the Bethlehem Thesaurus in SKOS

Since the Michael's Thesaurus is written in SKOS, and in order to integrate (or to map) the two thesauri, the Bethlehem Thesaurus must be also written (or exported) in SKOS. To this end, the Bethlehem thesaurus can be built using either the TMP which directly defines SKOS thesauri (figure 36) or the TMP2 based on ontoterminology using the SKOS interchange format (figure 37).

Figure 36: TMP platform

Figure 37: OTe for Thesaurus

4   Mapping between KYOTO and Arabic Ontology

As explained earlier (see section 2.2), KYOTO is a wiki-portal that provides a multilingual service to explore digital collections of environment and ecology objects and concepts. Because KYOTO supports SPARQL queries over structured data and using a rich ontology (unlike other portals that supports typical informational retrieval-based search) the Arabization of KYOTO fully depends on the mapping between the Arabic Ontology and KYOTO ontology.

The KYOTO ontology (see section 2.2) is a rich ontology that covers the very general and abstract concepts in a formal and consistent manner following rigorous and explicit criteria - based on the DOLCE-Lite-Plus (DLP, a top level ontology). As well be explained in the next sections, mapping between KYOTO and Arabic ontologies is an important step, not only for "Arabizing SPARQL- queries" in the KYOTO portal, but because KYOTO is also linked and mapped with several other ontologies and languages.

The mapping between KYOTO and Arabic ontologies was completed through producing a mapping between the top and core concepts of *Arabic Ontology* and KYOTO Ontology. BZU and BBAW conducted several meetings and discussions to achieve this goal. BZU and BBAW demonstrated their in-house tools and resources during the SIERA Kick-off Conference[28], where KYOTO and the Arabic Ontology project were presented and discusses. Moreover, in

---

[28] http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-kick-off-conference/. November, 2011.

the lexical semantic meeting in Trento[29] SIERA partners (BZU, BBAW and UNITN) focused on establishing this mapping and discussed how to extend KYOTO to Arabic. In addition, BZU and BBAW discussed and reevaluated the obtained mappings while participating in the 7th Global WordNet conference[30]. Along with many online discussions and emails the partners succeed in establishing and linking the Arabic Ontology Top Levels concepts to the KYOTO Ontology. Next, the mapping activities between AOTL and the KYOTO Ontology are provided.

4.1 Mapping between the Arabic and KYOTO Ontologies

The mapping between the Arabic Ontology and KYOTO Ontology concepts was done in two phases. In the first phase, BZU Sina mapped most abstract/top concepts about 63 concepts in the Arabic Ontology to the equivalent concepts in KYOTO, which were then reviewed by BBAW.

BZU Sina has performed a full and deep investigation of the mapped concepts and provided a detailed description and discussion for each mapping. Figure 4.1 illustrates the amount of details and discussions regarding the map between KYOTO and Arabic concepts -this example shows the Arabic concept "**Object/موجود**" and "**Endurant**" in KYOTO.

موجود	شيء له ذات مستقلة بنفسه وحاضر كليا في الزمن، يدرك بذاته (قياساً)، أو لذاته (اعتبارا).
Object	An entity that is wholly and independently present in time, and is realized either for its concrete or social existence.

Synonyms: ذات، كائن، شيء، قائم، حقيقي، واقعي، قيموم

Description: Objects are physical or social entities that we can point at and realize independently from any other entities. We look at a chair and realize its independent existence; the same goes for human beings, artifacts, and also geographical areas. Objects are independent entities that have no temporal parts; i.e. they persist over time maintaining their identity and have a location in time. This means that objects will always be wholly present in the times they're present in; for example, Jack will always be present, and will always be known and realized as Jack during his entire existence – life and after his death, BZU is wholly present in time, and will always be recognized as itself during its entire existence, my laptop is wholly present in time and will always be recognized as my laptop during its entire existence.

This class "Object" is equivalent to the class that's called "Endurant" in both DOLCE [2] and KYOTO [3], and is equivalent to the class "Independent Continuant" in BFO [1].

Objects have two types based on the way they're realized, either physically; i.e. physical objects, or socially; i.e. social objects, where there is no instance of an object that isn't either a physical object or a social object, and instances of physical objects cannot be social objects and vice versa.

Formally speaking:

$\forall x . PhyiscalObject(x) \rightarrow Object(x)$ *Every Physical Object is an Object*

$\forall x . SocialObject(x) \rightarrow Object(x)$ *Every Social Object is an Object*

$\forall x . PhysicalObject(x) \cap \forall x . SocialObject(x) = \emptyset$ *No instance of Physical object could be a Social Object, and vice versa.*

Identity Criteria:

We distinguish between two instances of object based on their presence as a whole in time. Thus, we consider two instances of Object to be exactly the same instance, if and only if they have the exact whole (physical of social presence) at a certain time. For example; if Jack is a physical object "i.e. a Natural person" that exists as a whole in time, if there existed another Natural person "P" at the exact time and place where Jack exists, then this Person is also Jack.

Also, let Birzeit University be an instance of a social object (i.e., as a legal person) that is present as a whole in time. If there existed another social object "S" that is we realize (i.e., recognize) to be exactly the same as Birzeit University in time, then this instance is also BZU.

Unity Criteria:

During their existence, Objects are treated as *wholes at this high level*, although they might consist of parts at the lower levels. During their existence, all instances of object at a high level; i.e. even it has two subclasses (social and physical) are generally treated as wholes

Even though some objects; i.e. physical objects, consist of parts; i.e. have parts, where changing these parts may affect the identity of these instances and they may become not instantiated, for its class, they still are represented as entities.

Rigidity:

Being an Object is Rigid (R+); instances of object will always be objects (that are wholly present in time of their presence and can either be realized for their concrete or social existence) during their entire existence. No instance of Object can stop being an object; i.e. stop being wholly present in time of its presence or realized for its concrete or social existence.

Instances: my house, My father's company, State of Palestine, Queen (the band), the statue of liberty, Jack's liver, Kitty the cat, H1N1 virus that affected Peter, an amount of water, Jupiter.

References:

[1] Masolo, Claudio. Borgo, Stefano. Gangemi, Aldo. Guarino, Nicola. Oltramari, Alessandro. WonderWeb Deliverable D18 - Ontology Library (final). Technical Report, 2003.

[2] Smith, Barry. Basic Formal Ontology 2. Draft Specification and User's Guide, Technical Report 2012.

[3] Hicks, Amanda. Herold, Axel. **KYOTO Ontology**: Foundations and Acquisition. Berlin-Brandenburgische Akademie der Wissenschaften. First Kyoto Workshop, Amsterdam, February 2–3 2009

{Thing}	A separate and self-contained entity

Figure 4.1: Example of the map between the Arabic and KYOTO Ontologies

The mapping focused on the comparisons and links of the Arabic Ontology with the KYOTO Ontology with a particular emphasis on the definitions and descriptions of the terms. 45 concepts subject to comparison between Arabic Ontology and KYOTO are extremely clear and accurate, where 18 classes subject to comparison stand in need of more clarification or revision. All over, about 72% of the comparisons are accurate. This outcome of mappings accuracy between the Arabic and KYOTO core concepts is particularly important for several reasons (see more in section 4.3). First, the upper concepts in an ontology are the most abstract and are frequently not lexicalized or intuitive to laypersons, so extra care is needed to ensure that the technical meaning of the most general terms are sufficiently taken into account during the mapping process. Second, because ontological hierarchies rely on transitivity of the subclass relation, errors at the top levels of the hierarchy will propagate downward. Any inaccurate mapping will strongly influence the mappings of the lower levels. Finally, since KYOTO is based on semantic web technology that utilizes inferences in queries, any inaccuracy in the mapping between the two ontologies may lead to invalid inferences that will affect the results of the query. This is unlike search engines where the irrelevancies of the retrieved results can be tolerated. The full mapping between KYOTO and the Arabic Ontology was reported in deliverable D2.2, but a revised version can be downloaded from:

https://www.dropbox.com/s/nnspzxjhjjziklo/ArabicOntologyTopLevels_Ver21.pdf?dl=0

4.2 Problematic Mappings Identified for Future Cooperation

In what follows we present the most problematic and mappings regarding specific comparisons and links of the Arabic Ontology AO with the KYOTO Ontology; these require longer investigation, and the SIERA partners agreed to continue resolving them after the project's end:

1. ***Abstract*** – *Abstract* in the AO is narrower than KYOTO's *abstract* since the definition in AO has more qualifications. At the same time, AO's *abstract* is also broader. In particular, *SpaceRegion*, *TimeInterval*, and *Quantity* are subclasses of *Abstract* in AO but not in KYOTO. Consequently, only some of the instances of AO *Abstract* are also instances of KYOTO *Abstract*, but all of the instances of KYOTO *Abstract* are also instances of AO *Abstract*.

2. ***Attribute*** – The KYOTO Ontology does not have a class by that name. However, KYOTO, following DOLCE and in contrast to BFO, does distinguish between qualities and their values. Attributes in AO are values of qualities that do not use units of measure. This is very similar to DOLCE's *region* with the major difference that KYOTO models the values of measurable qualities under *region* (cf. *definite quantity*, and *number*). Everything that is an instance of AO *Attribute* is also an instance of KYOTO *region*, but there are instances of KYOTO *region*, that are not instances of AO Attribute. To deal with such cases, SIERA partners agreed that each quality should have a scale, and the accurate mappings to be performed within such scales (see the point below).

3. ***Quality*** – "Every *Quality* node should have a *Region* (or, a bag of *Regions*)"; then the attributes are linked to the *Regions*. In other words, we defined *Regions* as the spaces where the *Attributes of Qualities* are located. For example, color of a banana is a quality that can change value over a period of time. The attributes are the specific values for this quality, green, yellow, brown, etc. These attributes are arranged in a color region that contains all and only the possible values or attributes for the quality color.[31] An outstanding question in the field is how to best model the relations among multiple regions and their parts. With a view to addressing some of these questions. the Partners discussed whether regions could be hierarchically represented. For example, Height, Width, and Depth are all kinds of Lengths, so it is tempting to model these using the subclass relations. Every height quality is a length quality, but it not clear whether every height region is a length region that is distinct from a width region. More specifically, we need to investigate the identity criteria and distinguishing characteristics of quality regions so that we may model regions, their parts, and attributes in a way that reflects

---

[31] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L. 2002. Sweetening Ontologies with DOLCE In A. Gómez-Pérez, V.R. Benjamins (eds.) Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Springer Verlag, pp. 166-181

their ontological properties and allows for correct reasoning regarding the identity and relations of qualities and attributes. The question is, how can we best formalize the relations among regions (e.g., as an equality, or part-of)? Answering this question is a necessary step to evaluating and validating the representation of qualities. These issues will be examined in the near future. To examine the hypothesis that "Every (*Quality*) node should have a *Region*". The partners designed an experiment behind the project goals that aims at defining the *Regions* for each *Quality* node in the AO, then linking them with their correspondence attributes. This experiment involves identifying and modeling regions for each quality and establishing the relations among the regions. Furthermore, some Quality-related issues should be further investigated, such as, *Mood* is a subclass of *Quality* not an emotional state. Also it was decided that *Dispositional* is a subclass of *Quality*. Concepts like *Back* and *Face*, which contain many parts or organs that share the same location or region that are collectively referred to with terms such as "face" or "back", can be classified under a node titled "Body Part" instead of "Collection of Organs".

4. **Physical Attribute** – This can be compared to DOLCE's and KYOTO's *physical-region* with the similar observations for Attribute made above. *Physical regions* in DOLCE and KYOTO are the values of qualities that only inhere in physical endurants and include values that utilize units such as *length*.

5. **Time and Space** – Although AO observes that their definition is more specific than KYOTO's, it is worth noting that the instances are similar. Since investigating time and space is a huge topic by its own, it was decided to continue investigating lower levels of Time and Space as a future cooperation between SIERA partners.

6. **State, Role, and Disposition** – Although KYOTO does not have a class *Dependent Entity*, it was not clear that we mean different things by *State*. Whether or not we mean the same thing can be addressed by the question, are AO states homeomeric stative? Likewise roles in KYOTO roles and dispositions are dependent entities, though there is no named class of dependent entities with this name. Their definitions are quite different, but it could be that their extensions are the same.

7. *Collection* – The DOLCE DNS definition is not correct. At any rate KYOTO does not consider collections to be containers. Also, some instances of AO *collection* are agentive and so do not match the KYOTO definition, e.g., a team of doctors.

8. *Social-agent* – This term is not the name of a class in KYOTO, so it is not clear which class is being referred to for purposes of comparison.

9. *Physical Object* – It should be noted that the class in DOLCE, and hence in KYOTO, is called *physical-endurant*. The parenthetical remark in the quoted definition is not in the original and makes a difference for the meaning.

10. **Material** – While DOLCE and KYOTO do not have a class called *Material*, this class is very similar to *amount-of-matter*.

11. **Artifact** – The term *material artifact* in KYOTO is inherited from DOLCE-DNS, and the AO also contains *artifact*, but this concept is still not accurately defined.

12. **Entity** – It should be noted that the root node of KYOTO (and DOLCE) is labeled "particular" rather than "entity". This is a minor point, and in so far as both are root nodes, they can be considered as comparable.

4.3 The usefulness and usage of the AO-KYOTO mappings

The majority of current comparisons of the Arabic ontology concepts with the KYOTO Ontology precisely deal with abstract and technical issues accurately, and this is a significant achievement. Partners agreed that the *Top Levels of the Arabic Ontology* contain a relatively large number of concepts and the amount of work and attention to detail required in developing a Top Level Ontology is vast. *Ontologies are improved by successive revisions*, and for a first release, Top Level Ontology of the Arabic ontology represents an impressive effort. Iterative efforts to improve both the Arabic Ontology and the KYOTO Ontology *in tandem* will be continued in future discussions and experiments.

It is worth noting that the provided mappings between the Arabic Ontology and KYOTO core concepts go together with task 2.2.3 "Establishing a Framework for Mapping Between WordNet and the Arabic Ontology". The BBAW has been carrying out a preliminary mapping of WordNet onto the KYOTO ontology that focuses on mapping the nouns of Core WordNet synsets to the KYOTO Middle Ontology[32]. All of the nouns from 1KCoreWordNet (aprrox. 600 sysets) are mapped to KYOTO Middle. These mappings are available for download at www.github.com/aellenhicks/KYOTO. Approximately 3000 nouns from the 5KCoreWordNet have received preliminary mappings. In future work we plan to compare our mappings with those recently produced by LOA from Core WordNet to DOLCE. The result of the mappings to KYOTO will be an indirect mapping of WordNet onto the Arabic Ontology via the KYOTO Ontology. Such mappings are crucial in enabling cross/multi-language applications such as machine translation and information retrieval[33]. In addition to being mapped to the English WordNet, the KYOTO Ontology has been mapped to Dutch, Spanish, Italian, Japanese, Basque, and Chinese in order to facilitate multi-lingual data extraction from corpora in each of

---

[32] Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). "Adding dense, weighted connections to WordNet." In: Proceedings of the Third Global WordNet Meeting, Jeju Island, Korea, January 2006.

[33] Vossen, P. (1998). Eurowordnet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht.

these languages.[34] The KYOTO system was designed to allow integration of further wordnets. The mapping of AO to the KYOTO Ontology provides a first step toward integration into the KYOTO architecture. Mapping AO and KYOTO also provides a first step to mapping AO to each of these six languages. The Multilingual Central Repository shows that mappings provide the potential to facilitate integration with large-scale knowledge bases such as DBPedia[35] and Geonames.[36] Of the seven languages previously mapped to the KYOTO Ontology, English, Spanish, and Basque are already in the MCR. Mapping AO to Spanish and Basque would, therefore, be low hanging fruit, and would allow the AO to take reap the benefits of multilingual mappings employed with the MCR.

For example the MCR takes advantage of semantic relations acquired from two corpora based on selctional preferences of the terms. These are encoded as noun-verb relations. Relations that are especially salient in one language can be utilized in the wordnets for another[37]. This will allow AO to benefit from the cross-lingual analysis conducted in the MEANING project. Such relations have also been shown to be useful in word sense disambiguation tasks[38].

---

[34] Vossen, Piek et al., KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures, In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), 2008.

[35] http://dbpedia.org/About

[36] http://www.geonames.org/about.html

[37] Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B. and Vossen P. The MEANING Multilingual Central Repository. In Proceedings of the Second International Global WordNet Conference (GWC'04). ISBN 80-210-3302-9. Brno, Czech Republic. January, 2004.

[38] E. Agirre and A. Soroa. Using the multilingual central repository for graph-based word sense disambiguation. In Proc. LREC 2008, Marrakech, Morocco, 2008. ELRA.

5 Arabizating and Extending OKKAM

The OKKAM Entity Name System (developed in the context of the OKKAM FP7 EU Funded project) is used to facilitate the disambiguation and linkage of Arabic entities with different languages[39]. The entities which were considered in this activity include people, organizations, places, and events. Such entities have different names (lexical labels) in different languages, which is a major challenge faced when integrating knowledge from different sources, cultures, and languages. SIERA partners have defined a set of activities in (task 2.2.4) that aim at resolving and linking Arabic entities with entities in multilingual portals to leverage of the knowledge across such multilingualism and culture diversities that the portals supports. In particular, resolving and linking Arabic entities to other Web entities across different languages through the OKKAM ENS.

Next, the OKKAMized Arabic entities datasets are introduced, and the Okkamization process of each of these datasets is presented. The activities carried out in Arabizing OKKAM ENS are also presented; these activities were introduced later in the project due to the unavailability of MICHAEL. The section concludes with the potential impacts of the performed activities and future joint works that agreed between the partners.

5.1 Enriching OKKAM with Arabic Entities Datasets

SIERA partners BZU, UNITN and CCHP (Centre for Cultural Heritage Preservation) in Bethlehem have conducted several meetings in order to demonstrate the in-house tools and requested data. The output of those meetings was the preparation of 1000 GIS-enabled cultural objects (Bethlehem historical buildings) to be included in OKKAM. The RDF version of the objects was produced by BZU and sent to OKKAM for integration. In addition, BZU provided an experimental sample of 31.000 named entities (about famous/old people) extracted from Arabic Wikipedia to be included in OKKAM.

Later in the second phase of the project, upon the project management board decision to Arabize OKKAM due to the unavailability of MICKAEL, BZU collected about 240.000 Arabic-English entity pairs from Wikipedia, the entities include name of people, organizations, and locations.

---

[39] P. Bouquet, H. Stoermer, and D. Giacomuzzi. OKKAM: Enabling a Web of Entities. In I3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.

5.2 The OKKAMization Process

The OKKAM team of the University of Trento extended the Entity Name System to support the identification of Arabic entities. Namely, a set of entities were OKKAMized. OKKAM makes the systematic reuse of globally unique identifiers possible by maintaining a globally unique and persistent identifier through the Okkam Entity Name System APIs[40]. Each identifier is tied with an entity profile, and a set of alternative identifiers. In particular, the entity profile[41] is used to support the execution of sophisticated entity matching algorithms, therefore enabling reuse of the entities' identifiers. In fact, once an identifier is defined in the Entity Name System, third parties can retrieve it by submitting identification requests to the Okkam Entity Name System search services. Currently, the services available are SOAP and REST APIs, and a Web Interface[42]. Using the latest, users can submit identification request using the Entity Identifier Request Language[5], and lookup the identifier of the entity of interest.

In the first phase of the project, two datasets of Arabic entities were OKKAMized; the dataset of Bethlehem's historical building and the Arabic entities for famous people extracted from Wikipedia. In practice, the Okkamization process requires the application of a semantic ETL (extract, transform and load) process. Namely, data were extracted from different sources (Wikipedia and Bethlehem CCHP GIS system) to produce RDF data using local vocabulary. Therefore, the first step was to load these data using the Open Refine tool, and apply several transformation functions on the data. The transformation can be categorized in three main areas:

a) Resolution of granularity problem. For example, splitting complex attribute to represent different atomic attributes, or aggregate atomic attributes to compose complex one.

b) Integration of contextual information. For example, the fact that the building is in Bethlehem, in Palestine.

c) Correction of errors and syntactical details.

Once the above content is produced, the provided attributes would be mapped towards the Okkam Identification Ontology to support the execution of Knowledge Based Entity Matching in the ENS; this step is called the Semantic Harmonization. Afterword, the entities are reconciled in the dataset of the ENS, and create new Entity Profiles together with a Global Persistent and Reusable identifier for the entities that were not already in the ENS. Moreover, details on the OKKAMized Arabic entity datasets are provided.

---

[40] http://api.okkam.org

[41] a set of attributes in the form of (key, value) pairs

[42] http://api.okkam.org/search/

First Dataset: The historical buildings of Bethlehem

The dataset was collected by BZU from the Centre for Cultural Heritage Preservation (CCHP) of Bethlehem, which produced a census of the historical building to be preserved. BZU Sina provided OKKAM with the dataset in RDF format (Figure 5.1 shows an excerpt).

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF          xmlns:ens="http://models.okkam.org/identification-ontology.owl#"
       xmlns:bb="http://www.cchp.ps/#"
       xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<rdf:Description rdf:about="http://www.okkam.org/eid-1c463018-c4d3-4b19-87ae-c298f2e58b53">
       <rdf:type rdf:resource="http://models.okkam.org/identification-ontology.owl#location"/>
       <ens:first_level_administrative_parent>West Bank</ens:first_level_administrative_parent>
       <ens:second_level_administrative_parent
xml:lang="en">Bethlehem</ens:second_level_administrative_parent>
       <ens:second_level_administrative_parent
xml:lang="ar">Bethlehem</ens:second_level_administrative_parent>
       <ens:timezone>GMT +3</ens:timezone>
       <ens:latitude>31.7029527901924</ens:latitude>
       <ens:longitude>35.2071602249924</ens:longitude>
       <ens:country xml:lang="en">Palestine</ens:country>
       <ens:country xml:lang="ar">فلسطين</ens:country>
       <bb:building_number>1</bb:building_number>
       <bb:parcel_number>0</bb:parcel_number>
       <bb:block_number>0</bb:block_number>
       <bb:number_of_inhabitants>55</bb:number_of_inhabitants>
       <bb:building_name>Building #1</bb:building_name>
       <bb:year_of_construction>1980</bb:year_of_construction>
       <bb:connection_to_sewage_network>0</bb:connection_to_sewage_network>
       <bb:availability_of_water_well>0</bb:availability_of_water_well>
       <bb:urgent_need_for_intervention>0</bb:urgent_need_for_intervention>
       <bb:additions xml:lang="en">No Additions</bb:additions>
       <bb:additions xml:lang="ar">لا يوجد إضافات</bb:additions>
</rdf:Description>
```

Figure 5.1: excerpt of the RDF

Around 1000 buildings' descriptions were provided, and 643 of them were OKKAMized. The partial OKKAMization of the dataset is due to the fact that some buildings provided very little number of attributes. This would in principle make the retrieval of their identifiers particularly complicated. Therefore, we decided to postpone the OKKAMization of the remainder of the historical buildings. The descriptions of the buildings in Bethlehem included geospatial information that allowed us to place them on the map of entities Okkam is maintaining. Figure 5.2 shows a view of such map with the detail of one of the buildings is presented in.
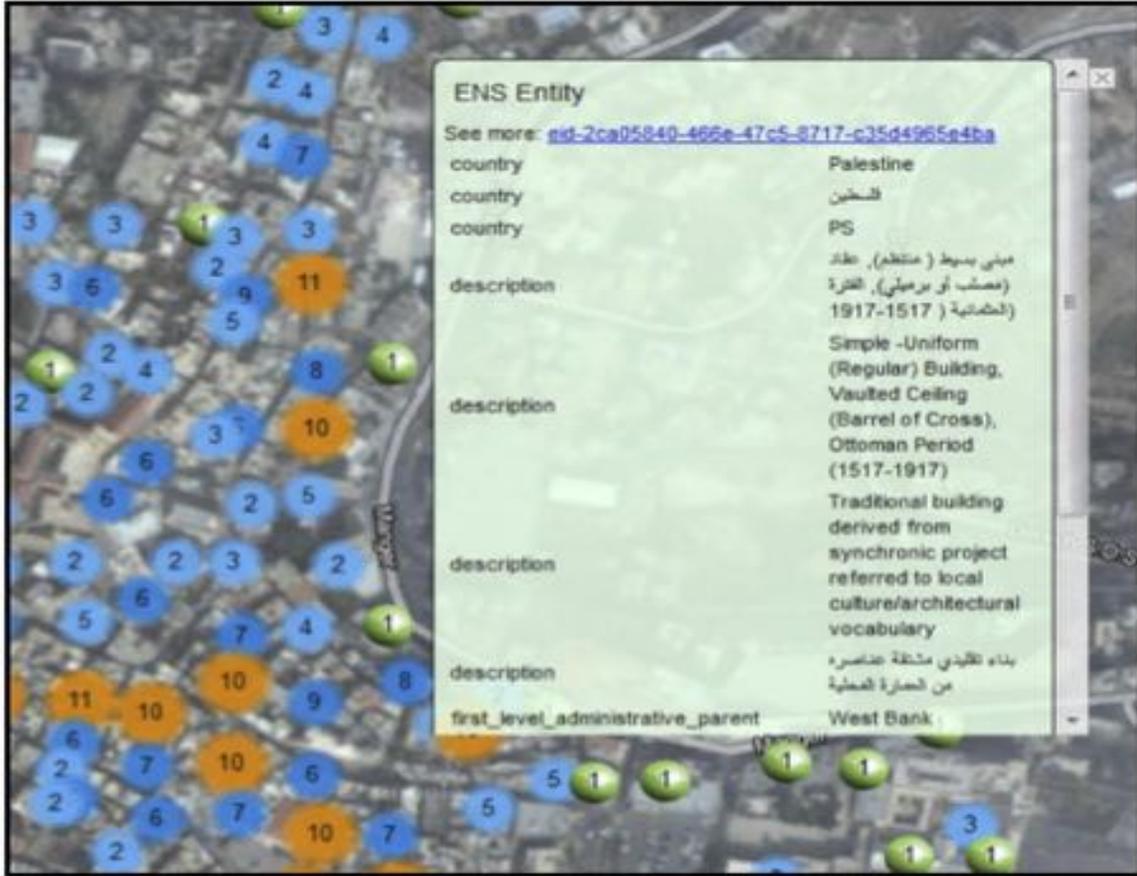
Figure 5.2: A view of the map of Bethlehem with the details of one of the building OKKAMized.

Second Dataset: Named entities extracted from Arabic Wikipedia

The Arabic Wikipedia named entities dataset was built crossing the descriptions produced by the BZU Sina team from the Arabic Wikipedia, DBpedia. This operation was done relying on an extention of the Open Refine tool made available by OKKAM. The DBpedia entities were integrated with the Arabic names of entities obtained processing the Arabic Wikipedia, to produce descriptions that could be OKKAMized. This allowed us to create 1107 new entity profiles integrating Arabic names for entities. This is just a fraction of the 30.000 extracted from Arabic Wikipedia.

The integration of Arabic names into the person profile allowed us to retrieve identifiers for such entities also using Arabic names, as showed in Figure 5.3 and 5.4. In fact, thanks to the extension of the descriptions with Arabic names, now entity identifies can be retrieved also using Arabic words.
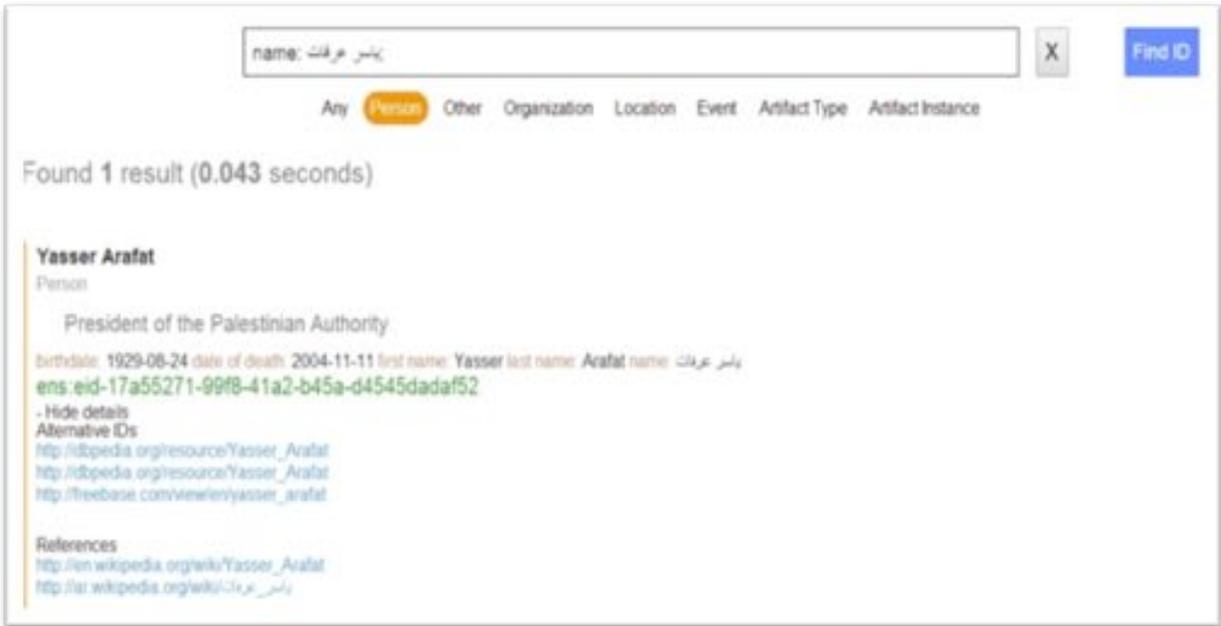
Figure 5.3: Screenshot of Entity Name System Lookup Interface searching using Arabic Names

Figure 5.4: Screenshot of a profile in the ENS Interface searching using Arabic Names.

5.3 Arabizing OKKAM Tools

As previously mentioned, and due to the lack of cooperation and accessibility of MICHAEL, the project management board has decided to include OKKAM as test bed to enable the investigation and integration of Arabic language and content into the OKKAM Name Entity System (ENS). In this regards, several online meetings and discussions have been conducted between the partners regarding the Arabic extension of the OKKAM ENS. Six main activities were planned to achieve the proposed goals; the activities and their main achievements within the reporting period are presented in the next sub sections:

5.3.1 Arabizing search: Integrating Arabic APIs into OKKAM

To enhance searching the Arabic entities in OKKAM ENS, the partners agreed to integrate Wojood into ENS. The following linguistic tools (Part of Wojood) were developed at BZU:

- *Arabic Language Detector*: A tool that returns true if the input is Arabic and false if the input is non-Arabic (non-Arabic includes: Urdu, Persian, English, etc.).
- *Arabic Spellchecking tool*: A tool that spell checks the input query (if the query is an Arabic one) and suggest (if misspelled) possible replacements.
- *Arabic Query Expansion*: Expands the input query (if Arabic) to introduce possible expansion for it. For example "in English": the word "study" will have an expansion list that includes: studies, studying, studied, etc. The English example is presented just to demonstrate the idea; the tool will be for Arabic words.
- *Arabic Light Stemming*: A tool that normalize the input text by removing some unnecessary prefixes and suffixes from Arabic input.

BZU Sina designed and finalized the above tools to work with minimal installation and provided them as a standalone service that is easily integrated with any search functionality built with Java programming language. UNITN re-packaged the tools using Apache Maven to fit their system requirements.

In a first phase, it was decided to integrate the Wojood API in the Entity Lookup Interface that was previously described. This enables final users in checking the results of the linguistic library, and fix issues related to misspelling and auto-complete functionality. In a second phase, OKKAM will integrate the query expansion function into the core query processor system of the ENS, to enhance recall of possible variants of Arabic names. In fact, the integration of the Arabic APIs ought to be part of a more complex refactorying of the ENS Core search engine. This work has been planned so that Arabic expansion can happen also when queries are submitted through web services calls, or third party applications. However, for the moment, users can play with Arabic language using the lookup interfaces. The ENS lookup interface

now enables suggestions on misspelling or autocomplete bases on the Wojood API as shown in the pictures below:

Searching for: Palestine

Searching for: abbas

Searching for: university
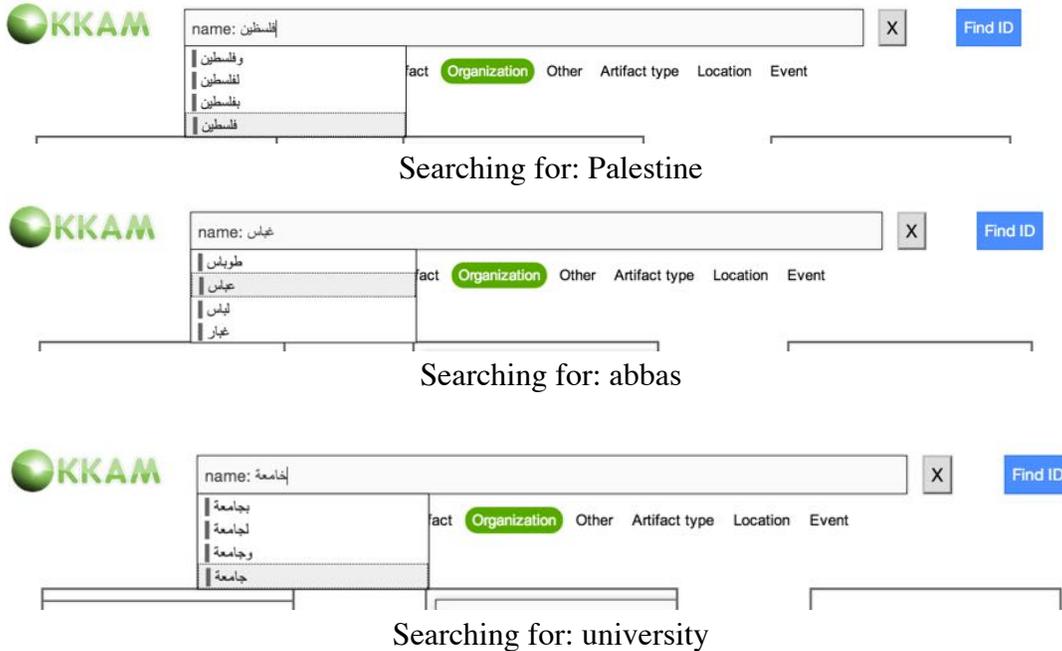
The Wojood API are integrated in the ENS Rich Web Application, as part the Java backing bean underpinning the search function. First, the library is triggered only if arabic language is detected. The detection is performed again using a specific function of the Wojood API. Then, while the user types the search terms, these are processed on-the-fly to produce the suggestions both in as autocomplete, or typos fixing. The APIs, once configured, are quite fast, and allow to perform these types of operation without causing any lag in the user lookup experience.

5.3.2 Arabizing OKKAM ENS search interface

BZU Sina with the cooperation of OKKAM-UNTIN has localized the web search interfaces of OKKAM NES (http://api.okkam.org/ and http://api.okkam.org/search/). Figure 5.5 and 7.6 shows the Arabic interface of the system.

Figure 5.5: the Arabization of the ENS home page
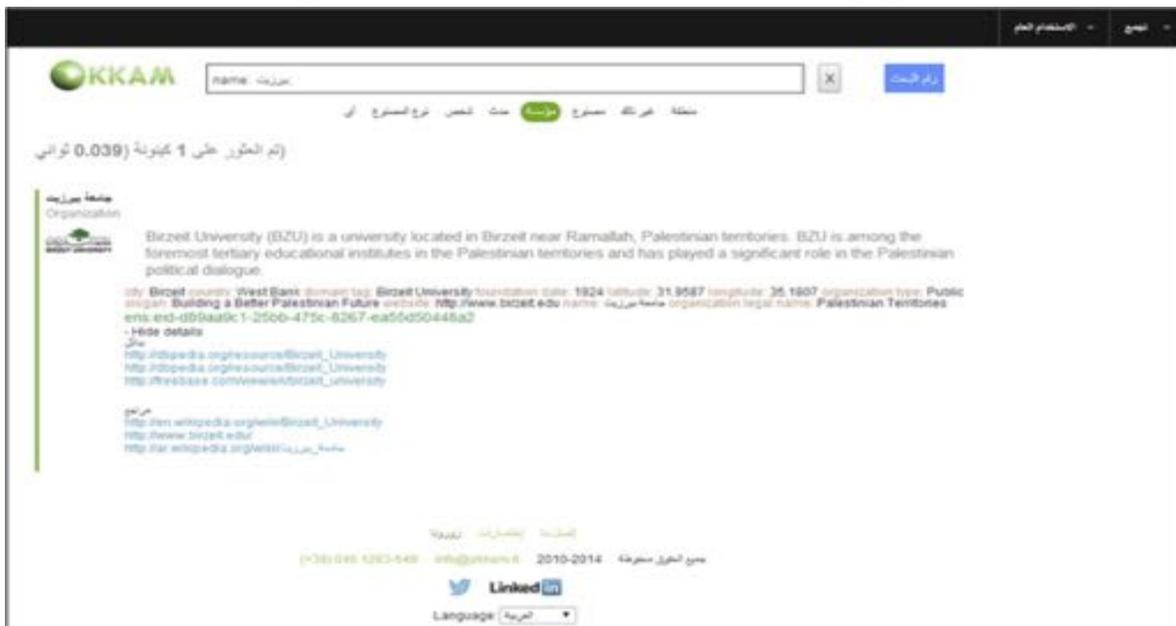
Figure 5.6: the Arabic interface of ENS

5.3.3 Localization of the ENS Ontology

To enable Arabic/multilingual searching functionalities the core ontology of the ENS was localized. BZU Sina has provided localization for the ENS ontology, which consists of 7 top-classes, each attached with a set of properties (in total 36 object properties, and 79 data properties). The arabization of the OKKAM ENS ontology can be found in the online ontology available at http://models.okkam.org/identification_ontology.owl.

Protégé (an ontology editing tools) was used to facilitate this task; Sina has provided the localization for all the ENS ontology's comments and labels. Figure 5.7 shows the localization

of the class "event" using Protégé. Figure 5.8 shows a sample of the ENS Ontology in OWL format (including the Arabic localization).
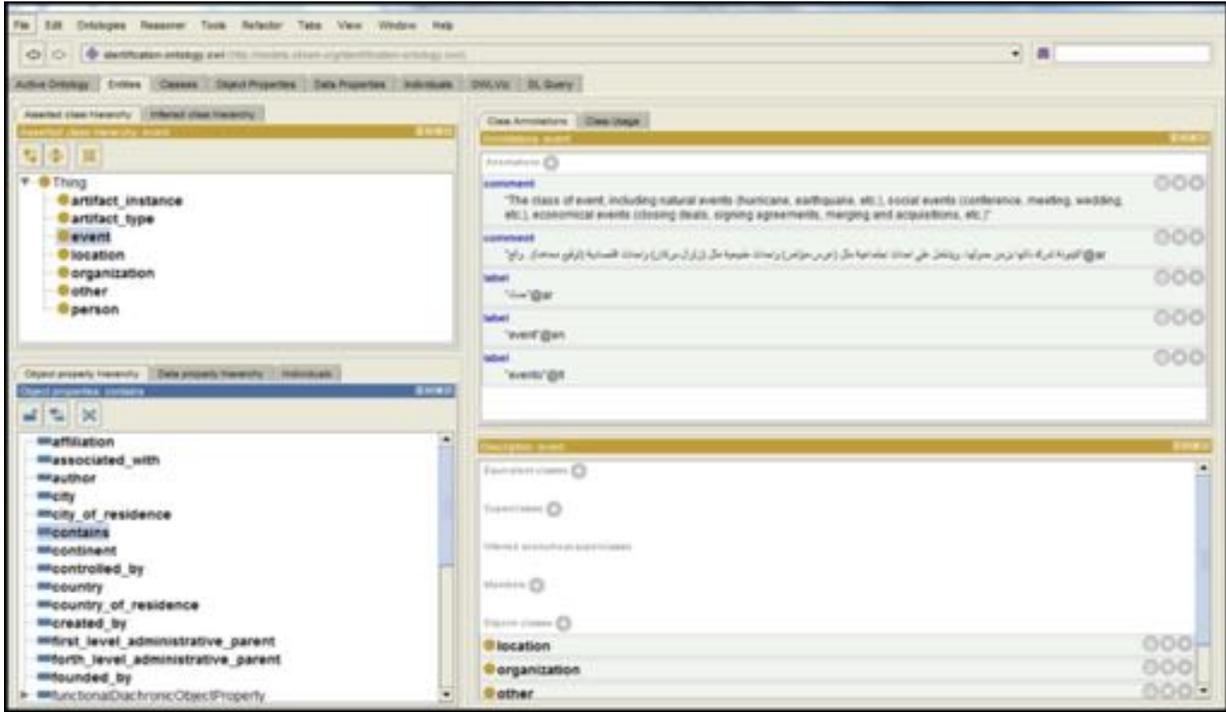
Figure 5.7: Localizing ENS using Protégé

Figure 5.8: An excerpt of the ENS OWL ontology the "event" class including the Arabic extensions

5.3.4 Extending ENS with Arabic Entities

BZU Sina has prepared a sample of more than 240,000 Arabic entities,which were extracted form Wikipedia. The sample as shown in the excerpt Figure 5.9 includes the Arabic and English titles, as well as the Wikipedia Interlingua (Arabic and English) links.

In their turn, UNITN has OKKAMized the collection and provided and OKKAM identifier for each Arabic entity that was already managed by the Entity Name System (around 31000), Figure 5.10 and Figure 5.11 show the results of querying ENS over the new OKKAMized entities "ابن خلدون, Ibn Hkldoun" which is highlighted in the red box. The partners have agreed to extend this collection so to identify (OKKAMize) more Arabic entities in a future work.

Figure 5.9: An excerpt of the Wikipedia Arabic entities

Figure 5.10: An excerpt of the OKKAMized entities

Figure 5.11: searching over the Arabic OKKAMized entities

Moreover, the partners agreed to link the OKKAMized entities with the Arabic Ontology. To do so, the partners defined a set of attributes that will enable this mapping. For instance, the Wikipedia-based type and category properties of the entities, and then the entities would be matched to the proper class in the Arabic Ontology. The expected result is not only to populate the Arabic Ontology with such great amount of entities, but an important aspect of this mapping is that it would support multilingual (including Arabic) applications such as machine translation and cross lingual information retrieval.

5.4 Arabic DBpedia Initiative

Staring from the OKKAMized entities datasets provided in the above activities, the partner planned an ambitious goal, to create Arabic DBpedia like dataset. The completion of this task is beyond the project goals and it would be addressed in future collaboration. However, the core set of an Arabic DBpedia has been created relying on a set of the 240.000 Arabic entities that BZU Sina extracted from Wikipedia. The partial set includes all Arabic entities that's related to people with their Arabic equivalent English name (same as article name in Wikipedia), Category (Actors, Scientists, Football players, etc.), Birth Date, Death Date and Place of Birth. This set was built by BZU Sina based on processed Wikipedia articles content. The set is around 31,000 Arabic entities. The set is used to be matched with Okkam and will serve as the core set of Arabic DBpedia which opens the door to future collaboration. Figure 5.12 shows a sample of this partial set.

	A	B	C	D	E	F
1	English_Name	Arabic_Name	Category	Birth_date	Death_date	Birth_Place
2	Elena Dementieva	إيلينا ديمنتييفا	tennis biography	15-Oct-81		Moscow, Russian SFSR, Soviet Union
3	Elena Georgescu	إيلينا جورجيسكو	sports person	4/10/1964		
4	Elena Gheorghe	إيلينا جورجي	musical artist	7/30/1985		
5	Elena Gómez	إيلينا غوميث سيرفيرا	gymnast	11/14/1985		Manacor
6	Elena Novikova-Belova	إيلينا نوفيكوفا بيلوفا	sports person	7/28/1947		Sovetskaya Gavan, Khabarovsk Kray, Russian SFSR
7	Kostas Martakis	كوستاس مرتاكيس	musical artist	5/25/1984		Athens, Greece
8	Kosuke Kitajima	كوسوكي كيتاجيما	swimmer	9/22/1982		Tokyo
9	Kotono Mitsuishi	كوتونو ميتسويشي	person	12/8/1967		Tokyo, Japan
10	Paul Hermann Müller	بول مولر	scientist	1899-1-12	10/12/1965	Olten, Solothurn, Switzerland
11	Paul Heyman	بول هيمان	professional wrestler	9/11/1965		Scarsdale, New York
12						

Figure 5.12: Sample of the people partial set (the core of the DBpedia)

5.5 The Impact of Arabic Entities OKKAMization

The Okkamization of Arabic entities has a great significance at both social and economic perspectives. With the OKKAMized entities one can enable the aggregation and mesh up of data around the entities. This is very useful in defining many applications bridging the linguistic gaps. For example, linking Arabic sources to English (EU) sources through entities could enable support for Arabic people in Europe by presenting them content in Arabic once the

identity of the object of interest is disambiguated (via OKKAMization). As a counterpart, both Arabic and non-Arabic people in Arabic countries could be enabled in accessing information about the entity of interest once the identity of this entity is disambiguated through the ENS. As a result, one can define many different types of services both for tourism as well as economical activities.

Moreover, the OKKAMization of the Arabic entities using both English and Arabic names would enable the reconciliation of any source mentioning it with identity of the entity managed in the ENS. That is, any source that reuses the OKKAM identifiers can be in principle aggregated, integrated or mashed up to produce any sort of added value service. For example, processing Arabic news, it would be possible to link Wikipedia articles about the mentioned person, that is, the Arabic news could be aggregated around mentioned entities. On the other hand, if we consider the historical buildings entities, one can use the same persistent identifier for those buildings to produce in an independent and decentralized fashion where different types of information could be aggregated according to the application needs. For example, the OKKAMized GIS points could be used to collect stories and other user generated content about the historical buildings and then propose them this content when the tourist is on place. **As well be presented in the next section, a success story has been achieved in the project, thanks to the Arabization of OKKAM (multimedia tourist guide),** which allows tourists through their mobile phones scanning the Okkam's ObjectLinks platform, where QR codes are placed at the entrance of these buildings in Bethlehem. Of course, all this is possible only with the strong partnership that joins state of the art know-how about semantic technologies of both Okkam and BZU Sina, with domains, linguistic and market knowledge of the Arabic world of Sina, and the Web of Entities vision with the whole technological stack developed by Okkam.

5.6 Future opportunities

The close cooperation between the partners Sina and Okkam has encouraged them to move further and to continue this scientific cooperation in the future. The partners discussing the idea of developing an Arabic DBpedia (which currently does not exists). By developing the Arabic DBpedia and linking it to the others datasets different services can be supported. For example, newspaper editors can integrate news with content coming from the Arabic DBpedia or other DBpedia. This interesting idea among many other topics would be considered in a joint research between Sina and Okkam in the future.

5.7 Success Story: SMART Multilingual Tourist Guide in Bethlehem

5.7.1 Overview

In the aim of enabling close and sustainable scientific cooperation between EU scientists and BZU Sina Institute and enhancing the institute's capacity in scientific cooperation, BZU Sina, together with the Centre for Cultural Heritage Preservation (CCHP) and University of Trento/OKKAM team, have deployed an innovative and pioneering electronic tourist guide for heritage sites in Bethlehem, using QR technologies through the OKKAM ObjectLinks framework.

This initiative is fully coherent with the objective of the SIERA project, as it involves: (i) integrating in-house resources and tools provided by project partners; (ii) providing Arabic support in multilingual knowledge sharing portals; (iii) resolving and linking entities which appear in different services or datasets and reorganizing contents and services around them.

The goal of the initiative is to enhance the experience of Palestine's visitors and valorize the cultural heritage of Bethlehem City, through an innovative use of smartphones and tablets.

While walking in the old city, tourists can scan a QR code with smartphones and automatically listen to an audio guidance in their own language, as well as watch video and read further description, about a certain heritage site.

The initiative targeted 8 locations of cultural and heritage significance in the city, including the Church of Nativity, Manger Square, and Star Street, among other significant locations.

It is worth noting that through OKKAM's ObjectLinks platform, the system allows for bundling several media data and content types, accessible through a single QR code. The developed framework allows for dynamic content editing, which provides for efficient content generation and enhanced monitoring and statistics capabilities. The system was praised by the local and international tourists visiting Bethlehem, quoting that the new system was simple and easy to navigate, while providing very rich and multilingual content.

5.7.2 Intended Objectives

The expected results of the project include: (a) effective valorisation of Bethlehem's and Palestine's cultural heritage (CH) assets; (b) perception of an innovative society which cares for its CH assets; (c) increased presence of Palestine and its CH assets at a global level; (d) innovative support to tourist guides, who can benefit from multimedia material as a support to their oral illustration; (e) reduced costs for maintenance of CH information (e.g. replacement of interpretation panels); (f) reduced costs for information updates; (g) establishment of an extensible infrastructure, which can be developed in the future without new significant investments.

These objectives were achieved by: (1) identifying 8 locations of heritage significance in Bethlehem, to valorise; (2) associating a collection of relevant multimedia resources (texts, audio files, videos, Apps) for each object/location as links; (3) making these resources immediately and easily accessible to the visitors in Bethlehem city through printing and deploying the QR codes at each corresponding location, to be scanned through smartphones and tablets via the conventional QR code scanning applications.

5.7.3 Perception of the Local Community

The initiative was carried out in close cooperation with the Palestinian Ministry of Tourism and Antiquities, in addition to CCHP. Dr. Hamdan Taha, the deputy minister expressed the ministry's interest at the initiative, and hoped to extend its coverage to span the different touristic places in Palestine. Dr Taha also indicated that the initiative goes along the ministry's plans to digitize and document the Palestinian narration through modern technologies, which should enrich the tourism industry in Palestine.

Moreover, Mrs. Nada Atrash, the Head of the Research and Training Unit at CCHP, indicated that the initiative stands as an innovative and simple facility to provide and spread information and guidance to the visitors of the city, enriching the tourist's overall experience. Mrs. Atrash indicated that the system has been used by hundreds of tourists already where the free internet connection provided by a local ISP facilitated the adoption of the system.

Deploying the system wasn't challenging for the local teams, given the fact that the system allows for the valorization of existing available resources and content, and presenting them through an innovative mean. ObjectLinks does not replace contents and services which were developed in the past (like videos, photos, texts, event feeds, etc.), but helps making them easily accessible to visitors through their smartphones in the right place and in right context. This increases the value of past investments and stimulates new developments. In other words, this project will be the starting platform for future research collaborations with EU scientists in the area of semantic data management, smart cities, integration of electronic (web) services.

Moreover, the system allows for effective and dynamic management of CH assets, where once an objectlink is created for a particular site, the associated contents presented to visitors can be

changed through the web-based interface (or through the platform's APIs) without the need of changing the physical QR codes or the links used in other web pages. This means that, for example, the QR codes at the Church of Nativity can be updated at any time by authorised personnel to keep up with new events, new materials, new advices for visitors.

5.7.4 Sample ObjecctLinks

The sample QR codes below represent the ObjectLinks developed for example sites in Bethlehem city. Follow the instructions below to get the content collected for each location

Church of Nativity	Manger Square	Star Street
powered by OKKAM	*powered by OKKAM*	*powered by OKKAM*

The Instructions to use the Guide

1. Make sure you have a QR reader application ("QReader" is recommended on both IOS and Android systems).
2. Scan a QR code with the QR reader application.
3. Choose your preferred media options from the list, which typically include, but not limited to: Audio guide recording for the corresponding site; Youtube video link; a webpage providing relevant text for each site; additional links to Facebook or other social media means, and potentially tweets list; contact a tour guide, to call a tour guide personnel to provide further assistance.

5.7.5 Publicity and News Coverage

To assure wider outreach for the initiative's goals and activities, especially with the fact that QR technologies in general are still relatively unfamiliar for the general public, Sina Institute worked on a publicity and dissemination campaign based on two pillars:

Preparing introductory brochures

Providing brief guidelines on how to use the system, and outlines the initiative's goals and vision. 2000 brochures have been distributed to the local travelling agents in Bethlehem city, along with some of the local shops at the city center to have better outreach. The brochure can be found below.

Coverage by local and international news outlets

A press release (in Arabic and English) has been disseminated to many local newspapers and online news outlets, which received high attention and reposting at local media in Palestine and in Europe, as well as at Social Media by the general public. The press release has further been published by the notable technology news platform in the MENA region, Wamda, which provided for regional attention to the initiative as well. A listing of all the outlets that reposted the news item can be found below:

English press release: http://sina.birzeit.edu/news-and-events/smart_guide/

Arabic press release: http://sina.birzeit.edu/news-and-events/innovative-tourist-guide-ar/

In the Media

Interview with Dr Jarrar

Comment on and Like this post!

SMART Multilingual Tourist Guide Deployed in Bethlehem

بيرزيت تطوّر دليل سياحي آلي لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

بيرزيت تطوّر دليل سياحي آلي لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

صحيفة القدس صفحة 22 - بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

صفحة 6   بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

بيرزيت' تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم'

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

بيرزيت تطوّر دليل سياحي آلي لأهم المواقع في بيت لحم

بيرزيت تطوّر دليل سياحي آلي لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

بالتعاون مع مركز حفظ التراث في مدينة المهد وجامعة «تورينتو» بإيطاليا
معهد ابن سينا في «بيرزيت» يطور دليلا سياحيا آليا لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلاسياحيا آليالأهمالمواقع في بيت لحم

بيرزيت تطوّر دليلاً سياحياً آلياً لأهم المواقع في بيت لحم

بيرزيت تطّور دليل سياحي آلي لأهم المواقع في بيت لحم

بيرزيت تطوّر دليلا سياحيا آليا لأهم المواقع في بيت لحم

Turismo: a Betlemme con guida elettronica creata in Trentino
Bethlehem electronic tourist guide created in Trentino
Turismo: a Betlemme adesso c'è guida elettronica da Trentino

A Betlemme con l'audioguida (del Trentino)

A Betlemme guida elettronica da Trentino

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Betlemme con Tablet e Smartphone, le visite elettroniche di Unitn

Visitare Betlemme con tablet e smartphone

Betlemme tecnologica

Alla scoperta di Betlemme direttamente con lo smartphone. Ecco come si fa

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Visitare Betlemme è più facile per chi ha uno smartphone

Un progetto trentino per semplificare il turismo a Betlemme

Palestina su Ipad e guida elettronica

Visitare Betlemme con tablet e Smartphone

Secondary References

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode – macitynet.it

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

A Betlemme guida elettronica

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

Bethlehem electronic tourist guide created in Trentino

Visitare Betlemme: dall'ateneo di Trento la guida che usa QRcode

6 Arabizing Organic.Edunet

In line with the goal of task 2.2.2 activities, the project management board has decided, after the impossibility of extending MICHEAL portal, and in addition to Arabizing OKKAM, to additionally Arabize the Organic.Edunet portal.

There are three main activities that were achieved regarding the Arabization of *Organic.lingua* portal: Localization of the user interface and messages and extending its underlying ontology, providing sample content in Arabic, and integration of Sina APIs to support Arabic search and retrieval, which described in the following subsections.

It is worth noting that, a strong cooperation between *Organic.lingua* project[43] partners and BZU Sina was established early while participating in writing the FP7 proposal project "agriMUSE" (Innovative Multilingual Content & Data Analytics Services for Agriculture), which was submitted under the call *FP7-ICT-2013-SME-DCA*. The aim of this project proposal was to use the experience of the ongoing *Organic.Lingua* project in the interest of developing multi-linguality services over a large-scale metadata aggregation infrastructure and that will (i) incorporate a set of language technology components and (ii) allow multilingual portals and services to be created on top of it. Although the project was rejected, but BZU Sina has gained the experiences and the partnership with the proposal participation, which showed one of its advantages in this activity as well as future joint proposals.

6.1 Arabizing and extending Organic.Lingua Ontology

BZU Sina provided an Arabic localization for the *Organic.Lingua* Ontology, which is used as background knowledge for the *Organic.Edunet* portal search engine. This localization is considered a core step to enable and support Arabic searching functionalities in the *Organic.Edunet* portal.

In particular, BZU Sina has Arabized 291 concepts out of 381 concepts in *Organic.Lingua* ontology from English to Arabic. BZU Sina has also added an Arabic description for 32 concepts out of the 291 arabized concepts. BZU Sina will continue in arabizing the *Organic.Lingua* ontology concepts, as well as it will provide an Arabic description for the arabized concepts. The list of the Organic.Lingua ontology concepts[44] with their Arabization can be downloaded here: https://www.dropbox.com/s/q95f4hd3o3wfq29/Organic.LinguaOntologyArabized.docx?dl=0

The arabization activity was carried on using *MoKi* (the Modeling Wiki for ontology evolution). Figure 6.1, shows a snapshot of the English concepts on the left, and the

---

43 http://www.organic-lingua.eu/
44 https://dkmtools.fbk.eu/moki/multilingual/ organiclingua/index.php/Special:ListConcepts

correspondence Arabic translation on the right. Each concept is attached with a description[45]. Users (i.e., Ontology editors) can click the edit button and provide the appropriate translation (see Figure 6.2). Once the user completes his/her translation tasks, *MoKi* administrator has to validate his work and confirm the added content. Moreover BZU Sina has also provided the localization of the portal interface. Figure 6.3 shows the Arabic interface version of the Organic.Edunet portal.
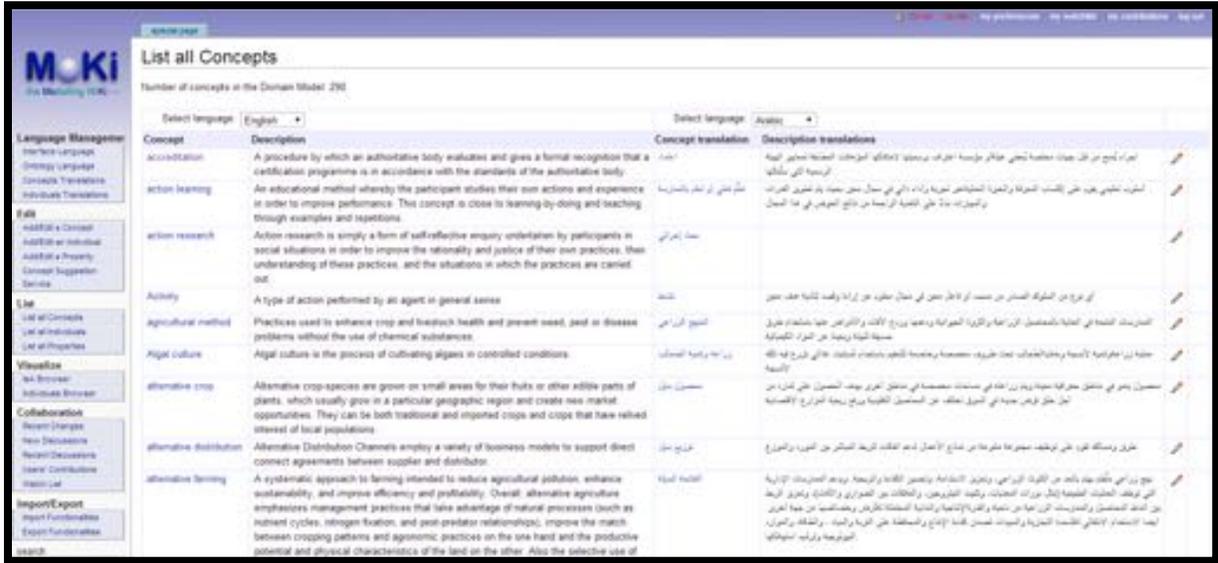
Figure 6.1: A snapshot of *MoKi* showing a list of English concepts and their correspondence Arabic concepts were each concept is attached with a description.

Figure 6.2: The user interface for translation.

Figure 6.3: the Arabic interface of the organoc.edunet portal

6.2 Enriching Organic.Lingua with Arabic

The second activity aimed to provide an Arabic content in the agriculture field in order to test as well as to demonstrate searching and consuming Arabic content on a multilingual knowledge sharing portal. BZU Sina has collected, edited, and uploaded lots of documents in Arabic (including documents, webpages (links), videos, PDF, images, etc). This content was uplaoded using the *Agricultural Learning Repository tool*[46]. Figure 6.4 shows a partial list of the provided content. The complete list of the provided content can be found in *the table below.*

---

[46] http://aglr.agroknow.gr
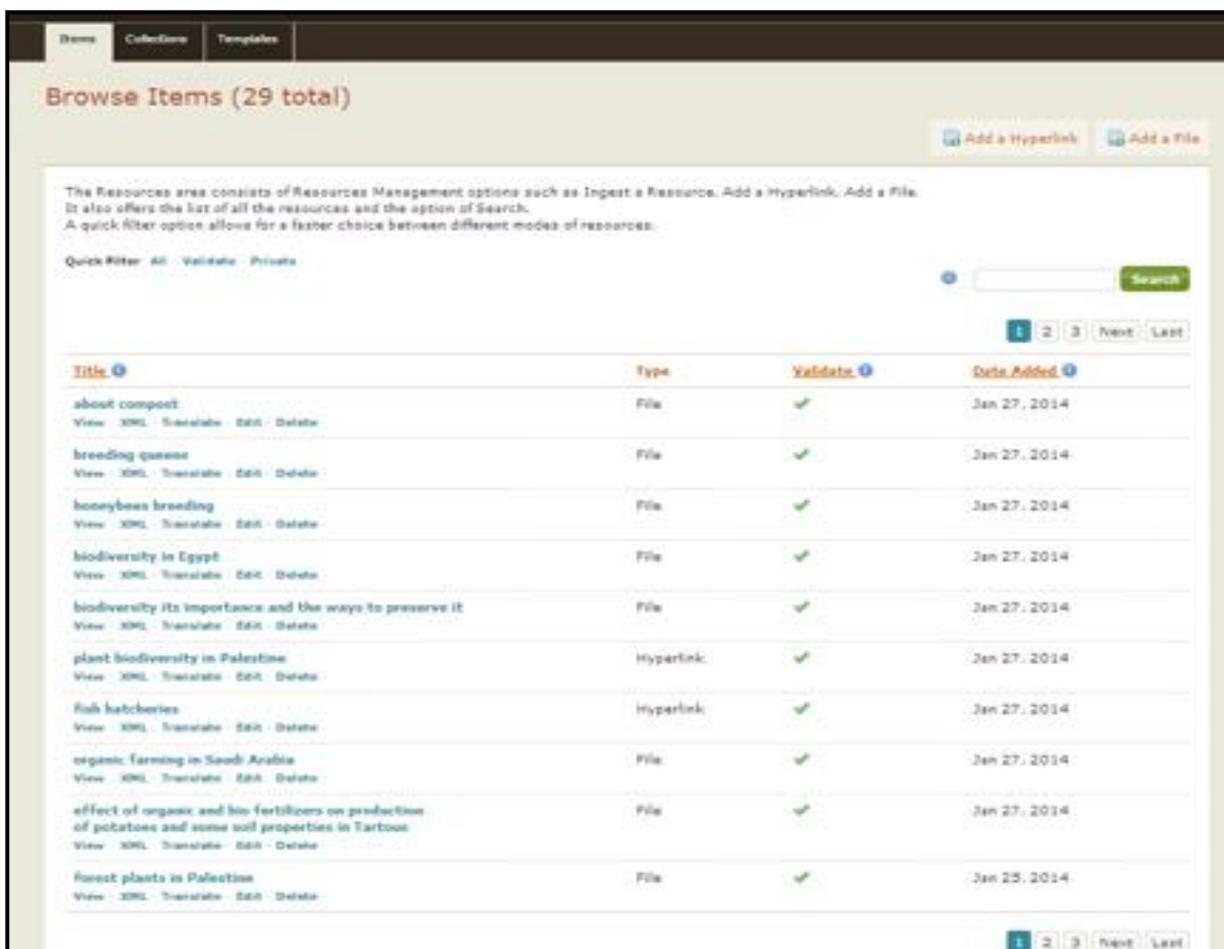
Figure 6.4: sample of the Arabic Content

Table: List of the 29 learning items uploaded through the *Agricultural Learning Repository tool*

No.	Title	Type
1	about compost *http://aglr.agroknow.gr/organic-edunet/archive/files/6da1dab751568d0f50ec4e8f258b46de.pdf*	File
2	breeding queens *http://aglr.agroknow.gr/organic-edunet/archive/files/bd5e9b932c7614de48d800e340a47da4.pdf*	File
3	honeybees breeding *http://aglr.agroknow.gr/organic-edunet/archive/files/b8932ed895049394b94a38b631b37ce5.pdf*	File
4	biodiversity in Egypt *http://aglr.agroknow.gr/organic-edunet/archive/files/d904f92f25f33a03613dc78e4697645a.pdf*	File
5	biodiversity its importance and the ways to preserve it *http://aglr.agroknow.gr/organic-edunet/archive/files/1caf3c9c35364bad410d314bd6580d34.pdf*	File
6	plant biodiversity in Palestine *http://www.youtube.com/watch?v=FBo50it2Kxo*	Hyperlink

7	fish hatcheries *http://www.youtube.com/watch?v=WLrV-9hH8Ms*	Hyperlink
8	organic farming in Saudi Arabia *http://aglr.agroknow.gr/organic-edunet/archive/files/90b4e4406b38b31204474ddec5fc3a68.pdf*	File
9	effect of organic and bio fertilizers on production of potatoes and some soil properties in Tartous *http://aglr.agroknow.gr/organic-edunet/archive/files/abde5872f24b342c43bc357acb4f59e7.pdf*	File
10	forest plants in Palestine *http://aglr.agroknow.gr/organic-edunet/archive/files/0ee07b75c6ca7484e1f824d4e37131d0.pdf*	File
11	biodiversity conservation *http://aglr.agroknow.gr/organic-edunet/archive/files/d0a2620b51fe7ad027f51aad743b0faf.pdf* **	File
12	the effect of organic fertilizer on wheat production *http://aglr.agroknow.gr/organic-edunet/archive/files/323920cda07bfdf53375d11f6f90d520.pdf*	File
13	biocontrol in Oman *http://www.youtube.com/watch?v=iMXyuGVCXOs*	Hyperlink
14	integrated pest management *http://aglr.agroknow.gr/organic-edunet/archive/files/eed7889f49aa70ce4a9f3581f2f06136.pdf*	
15	non-chemical control of olive pests and diseases *http://aglr.agroknow.gr/organic-edunet/archive/files/397304205d400045623234095964730e.pdf*	File
16	natural and environmental protection from agricultural pests http://aglr.agroknow.gr/organic-edunet/archive/files/d36378f9078eaef99ee1db29e21d1ae9.pdf	File
17	value-chain analysis of Egyptian aquaculture *http://aglr.agroknow.gr/organic-edunet/archive/files/437670a405619db2b583467360f373ae.pdf*	File
18	insect pests management *http://aglr.agroknow.gr/organic-edunet/archive/files/b7f68b248671239dff7715c1f26c30a1.pdf*	File
19	Organic and natural control of agricultural pests and diseases *http://aglr.agroknow.gr/organic-edunet/archive/files/1e1d280ce97ea21d7ec5633b03103318.pdf*	File
20	natural balanced fertilization *http://aglr.agroknow.gr/organic-edunet/archive/files/014ba87e111dc877591a7406068bff8f.pdf*	File
21	the development of organic farming *http://aglr.agroknow.gr/organic-edunet/archive/files/e3de7a24862189a3d482f8fac6bf1dad.pdf*	File
22	modern technical innovations in aquaculture *http://aglr.agroknow.gr/organic-edunet/archive/files/a944d0340a65bd677e0d432f98fdc26b.pdf*	File

23	aquaculture and fish farming in the Palestinian territories *http://aglr.agroknow.gr/organic-edunet/archive/files/65b13889730846f9faa11ffb67305e96.pdf*	File
24	fish farming (aquaculture) in deserts *http://aglr.agroknow.gr/organic-edunet/archive/files/4b702c0c50d5b30489c1b15a40e25e52.pdf*	File
25	diagnosis of bee diseases *http://aglr.agroknow.gr/organic-edunet/archive/files/f4e45b5eeee0c8a94546a18622755cf5.pdf*	File
26	how to make compost *http://www.youtube.com/watch?v=1nqdzuwzQVQ*	Hyperlink
27	honey production *http://www.youtube.com/watch?v=pS82ZnF9iC8*	Hyperlink
28	breeding queens *http://www.youtube.com/watch?v=wjvEyUk3gxM*	Hyperlink
29	agriculture in north iraq *http://www.aljazeera.net/programs/pages/c11906ea-6034-4af0-b700-fea9462034e2*	Hyperlink

To enable multilingual and multi-cultural sharing of the uploaded content, BZU Sina has annotated the uploaded content with relevant metadata. This step is important because it will enable the search over the Arabic content using the different languages that the portal supports.

For instance, the general information that describes the uploaded item was added (e.g., title, description, keywords). Moreover, the particular classification system and the collection under which the provided item is classified were provided. Figures 6.5, Figure 6.6 and Figure 6.7 show the main attached metadata which have been provided in both languages (English and Arabic); for a PDF version of the document in Arabic (forest plants in Palestine) which can be accessed here:

http://aglr.agroknow.gr/organic-edunet/archive/files/0ee07b75c6ca7484e1f824d4e37131d0.pdf.

Figure 6.5: General information metadata attached to the uploaded item
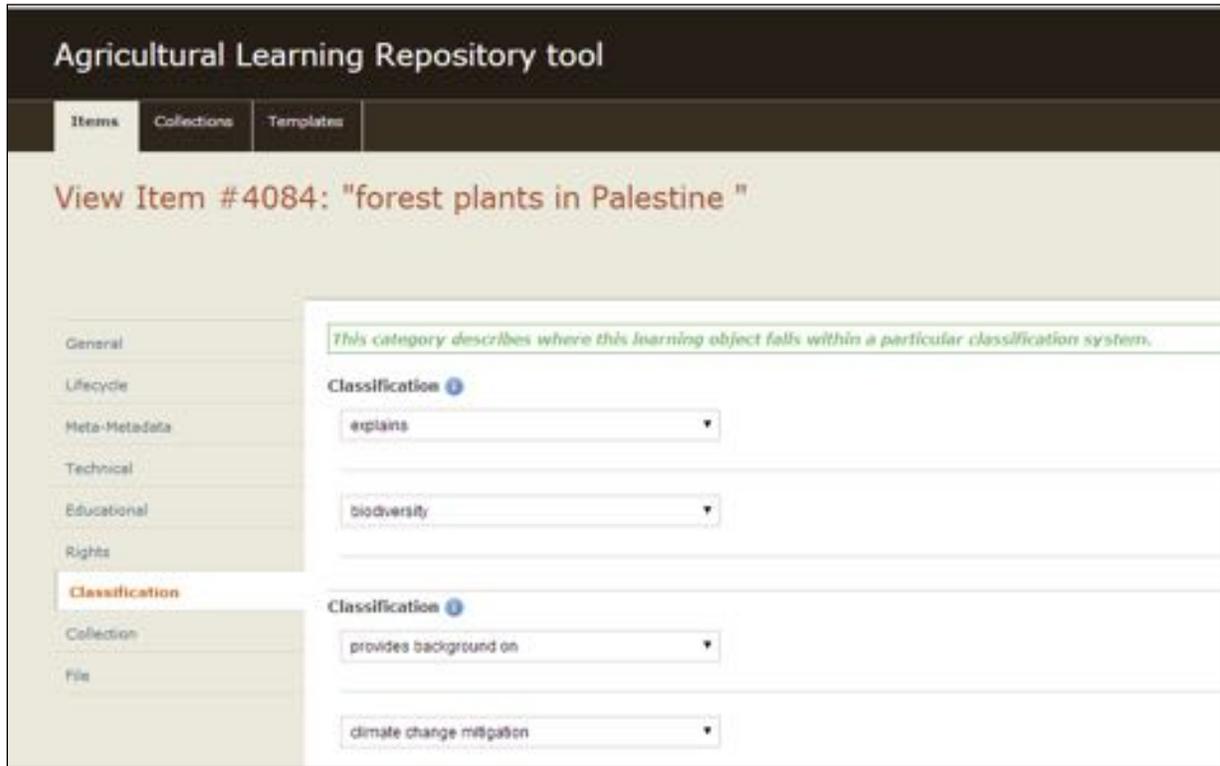
Figure 6.6: a classification of the uploaded item

Figure 6.7: define the collection of the uploaded item.

6.3 Arabizing Organic.EduNet Search: Integrating Arabic APIs

Similar to the Arabization of the OKKAM searching functionalities, BZU Sina tools and APIs were integrated with the Organic.Edunet portal to enhance searching the Arabic entities. The following NLP components (Part of Wojood, developed at BZU) were integrated:

- *Arabic Language Detector*: A tool that returns true if the input is Arabic and false if the input is non-Arabic (non-Arabic includes: Urdu, Persian, English, etc.). This tools is specially impotent as the Organic.Edunet portal supports many language, so that users can type in their preference and the portal will automatically detect their language.

- *Arabic Spellchecking tool*: A tool that spell checks the input query (if the query is an Arabic one) and suggest (if misspelled) possible replacements. This tool is used to help users autocorrect their misspelled quires.

- *Arabic Query Expansion*: Expands the input query (if Arabic) to introduce possible expansion for it. For example "in English": the word "study" will have an expansion list that includes: studies, studying, studied, etc. The English example is presented just to demonstrate the idea; the tool will be for Arabic words.

- *Arabic Light Stemming*: A tool that normalize the input text by removing some unnecessary prefixes and suffixes from Arabic input.

- *Arabic POS tagger Web Service:* a web service based on the usage of Stanford Arabic POS tagger. The API accepts the Arabic words as a string input, and at the server side then split the string into words, then returns the part of speech of each word as a JSON format to the client side. The service is hosted on nlpsina.appspot.com.

- *Bilingual dictionary Web Service:* a web services hosted at BZU Sina to provided an access to its multilingual dictionary -about 200 thousand translation pairs are provided through the service. Since Organic.Edunet is based on Machine translation, the service is used to return all the possible translations of the requested word either in English or Arabic as a JSON format to the client side. Figure 6.8 shows the (JSON) output of the translation of the Arabic word "سيارة" to the set of synonym English words {automobile, car, motor, motor car, sedan}, through the query sina.birzeit.edu/mabuhelou/dic/query_json_back.php?lang=ar&word=سيارة&callback=jsonSinaBiDicApi
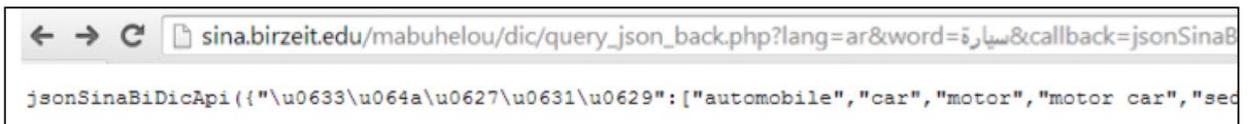
Figure 6.8: Arabic Bilingual Web Service out put for the query

Figure 6.9 shows the result of searching the Arabic contents in Organic.Edunet. The service works as follows: if the service intercepts an Arabic query it translates to English via the tools and APIs described above and if the translation is successful, the resulting English translation is then translated "normally" with the CLIR, then the relevant content is retrieved.
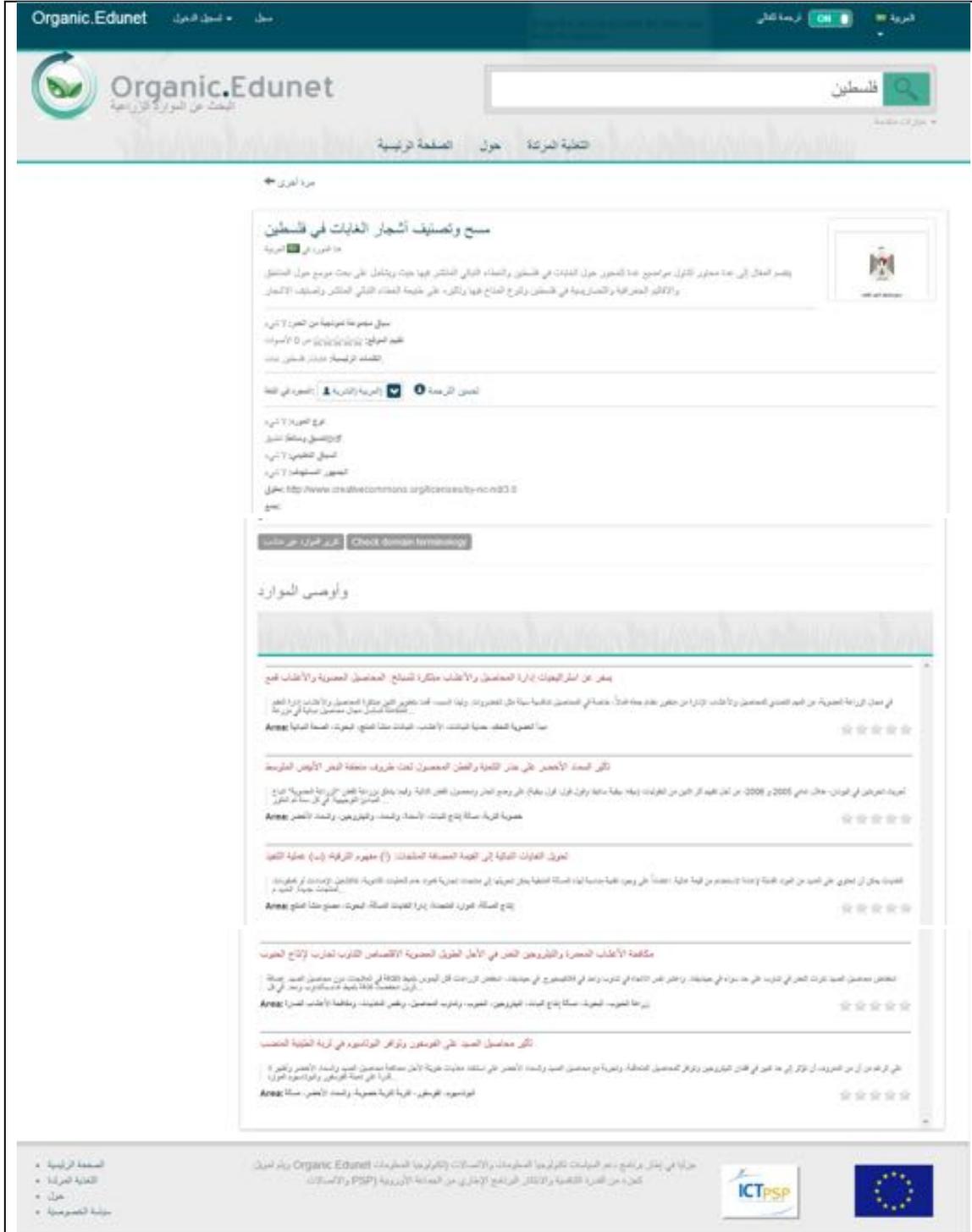
Figure 6.9: searching Arabic content on Organic.Edunet

7 Mapping Framework between WordNet and Arabic Ontology[47]

7.1 Introduction

This section reports on the implemented activities that are related to WP2, namely; (task 2.2.3) *establishing a framework for mapping between WordNet and Arabic Ontology*. The aim of task 2.2.3 activities is to establish a framework that links concepts lexicalized across different natural languages; in particular, to map Arabic concepts in the Arabic Ontology[48][Jarrar, 2011, 2012] to their correspondences English concepts in WordNet[49].

Wordnets (or, linguistic ontologies [Hirst 2004]) are semantic networks covering most common concepts in a natural language and provide knowledge structured on lexical items (words) of a language by relating them according to their meanings (concepts). WordNet [Fellbaum 1998] is the first and largest wordnet and is described in detail in Section 2.5. The success of WordNet motivated the construction of similarly structured lexicons for individual and multiple languages (multi-language lexicons). Recently, wordnets for many languages have been constructed under the guidelines of Global WordNet Association[50], which aims to coordinate the production and linking of wordnets.

The Arabic Ontology, is an ongoing project Sina-BZU, which has been described in detail in Section 2.6. At Sina-BZU hundreds of lexical resources (e.g., dictionaries) have been digitized and integrated into one lexical database. This database provides a good source for Arabic concepts, but lack semantic relations among the concepts. We argue that, by mapping such Arabic concepts into their conceptually equivalences in WordNet, one can (automatically) infer the relations among the Arabic concepts from the relations among the English concepts. The resultant relations are expected to provide an initial set of relations that can be manually validated and corrected.

However, mapping concepts lexicalized in different languages is a challenging task [Gracia et al., 2012]. The main objectives of the conducted activities in this task is to investigate a mapping framework that will be used in mapping concepts across different languages, in particular considering the scenario of mapping Arabic concepts to their equivalent English concept in WordNet; this mapping framework includes:

- the representation of the mappings,
- a theoretical interpretation of the meaning of these mappings,
- a description of a mapping algorithm and its components.

---

[47] This report contains unpublished work.
[48] http://sina.birzeit.edu/ArabicOntology/
[49] http://wordnet.princeton.edu/
[50] http://globalwordnet.org/

We started from approaches defined in the context of mono-language ontology matching and we extended them to cross-language ontology matching [Spohr et al. 2011] by considering the concepts' lexicalization. Cross-language ontology matching techniques, e.g., [Fu et al. 2012, Spohr et al. 2011], can play a crucial role in bootstrapping the creation of large linguistic ontologies and, for analogous reasons, in enriching existent ontologies. We also remark that the above considerations do not apply to the Arabic ontology only, but our definitions and approach are general and can be reused for other languages. In this regard, to achieve task 2.2.3 goals several activities were implemented by SIERA partners, the main activities and their results can be summarized as follow:

- SIERA partners demonstrated their in-house tools in details during SIERA Kick-off Conference[51]. This followed by several online discussions, emails, and meetings which were collocated with SIERA events (e.g., lexical meetings in Trento[52], and the 7th GWC).
- The partners succeed to establish and to define a mapping framework that map the Arabic Ontology to WordNet.
- Partners BZU, UniMiB, and BBAW have organized a discussion panel[53] on "Cross-Lingual Mapping Semantics" at the 7th GWC, in Tartu, Estonia in January 25th, 2014.
- A co-authored paper titled "Towards Building Lexical Ontology via Cross-Language Matching" [Abu Helou et al, 2014a] was presented at 7th GWC'14. The paper presents a classification-based mapping for cross-language ontology mapping, and defined an evaluation experimental for cross-language mapping semantics, through which a gold standard dataset can be constructed.
- The framework was used for mapping 10,000 Arabic concepts to WordNet.
- A cross-language mapping algorithm was also defined. The algorithm semi-automates the mapping process; by providing the users with the most appropriate candidate mappings. The algorithm and a preliminary evaluation were presented in a co-authored paper [Abu Helou et al, 2014b] submitted to OM'14 –ISWC'14[54].

The rest of the section describes the components of the mapping framework and is structured as follows. *Section 7.2* overviews the related work, and provides preliminary definitions used in the domain. *Section 7.3* presents a classification-based semantics for cross-language ontology mappings. We also show that our semantics can be used, in principle, to support the definition of cross-language ontology alignments by leveraging classification tasks performed by bilingual speakers. This method can be adopted to build reference alignments (gold standards) to evaluate automatic ontology matching methods. *Section 7.4* illustrates a cross-language mapping algorithm. We also evaluate the algorithm with a preliminary set of large-scale experiments, which helps to better understand the cross-language ontology matching problem and points to future research directions. In *section 5* we conclude with potential future research directions.

---

[51]http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-kick-off-conference/.
[52]http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-wp2-meeting-in-trento-italy/.
[53]http://gwc2014.ut.ee/index.php?v=panel
[54] http://om2014.ontologymatching.org/

7.2 Related Work and Background Definitions

Manually establishing mappings between two lexical resources of any kind (e.g., set of concepts, linguistic ontologies, or ontologies) is often unfeasible due to the size of the resources to consider and to the available resources [Jarrar, 2006, 2011, 2014]. For this reason, *ontology matching* was proposed as a.research field to study methods to automatically establish semantic relationships (correspondences) between two (or more) ontology entities [Bouquet et al. 2003, Euzenat and Shvaiko, 2007]. For a general introduction to the Ontology Matching problem we refer to Section 3.2.5, where Ontology Matching is introduced to describe an approach to map two different thesauri.

In this section we focus on ontology matching methods proposed for matching ontology entities lexicalized in different languages. We summarize the detailed analysis reported in D22 and recent contributions (not acknowledged in deliverable D2.2). We focus on computational approaches proposed to map concepts lexicalized in different languages and on the interpretation of the established mappings. Then, we introduce formal definitions for the ontology and the ontology matching problem. We also define the structure and the semantics of mappings in the mono-lingual ontology matching. Then we give an overview of the cross-language and multi-language ontology matching in relation to the mono-language ontology matching definitions, considering the ontology lexicalization. Finally we overview the classification-based approach for the weighted mappings proposed for mono-lingual ontology matching settings.

7.2.1 Cross-language Ontology Matching

The majority of the proposed matching techniques in these systems have mainly focused on mapping between ontologies that are expressed (lexicalized) in the same natural language (so-called, *Mono-Language Ontology Matching*, MOM) [Shvaiko and Euzenat,2013]. However, in the last years, notable efforts [Spohr et al. 2011, Fu et al. 2012] were made in order to overcome, the language barriers; the problem of matching two ontologies that use more than one language each, at the same time they share (at least one) the same languages (so-called, *Multi-Language Ontology Matching*, MLOM). A specific case is when the two ontologies do not share any languages to be matched (so-called, *Cross-Language Ontology Matching*, CLOM) [Spohr et al.,2011].

In general, *translation-based* approach [Fu et al., 2012] is considered as a solution to resolve the cross-lingual issue, to transform the cross-lingual problem into a mono-lingual ontology matching one. That is, to translate the ontological elements (e.g., concepts, comments, etc) of one ontology in the language adopted by the other ontology using automatic machine translation tools. Translation based matching systems leverages on machine readable dictionaries [Nagi et al.2002, Liang et al, 2006], and machine translation tools [Spohr et al, 2011; Fu et al.2012, Trojahn et al.2010]. However, the quality of machine translation systems is limited and depends greatly on the pair of languages considered [Spohr et al.2011]. Moreover, such approaches depend on the structure of the ontologies which will be matched (i.e., mapping between existing ontologies). Another interesting works for resolving the cross-lingual issue

exploit Wikipedia [Hassan and Mihalcea 2009, Hertling and Paulheim 2012], such approaches used Wikipedia inter-lingual links as a resource of multilingual lexicalization [Jarrar, 2014]. However, such approaches are limited and depend on the lexical coverage which are provided by Wikipedia inter-lingual links. A remarkable approach that disambiguates and discovers the proper semantic (sense) of keywords, more than just exploit machine or dictionary translations was presented in BabelNet [Navilgi and Ponzetto 2012]. Navilgi and Ponzetto combined the Wikipedia multilingual knowledge (inter-lingual links in Wikipedia) with machine translation to determine the appropriate senses among set of translated candidate mappings. *Appendix A* provides a comprehensive analytical comparison of the recent ontology matching systems.

A semi-automatic mapping framework was proposed to map concepts (synsets) in different languages by combining translation tools and word sense disambiguation into a hybrid task.

In the work carried out in the SIERA project we defined a mapping algorithm for constructing linguistic ontologies inspired form crowdsourcing framework presented in [Venetis et al. 2012]. We see the problem of mapping unstructured concepts to structured one, as a maximization problem that retrieve top-k mappings from a set of sorted candidate mappings. The algorithm can be viewed into two folds; translation and sense selection (disambiguation) tasks, details on the mapping algorithm and a pre-evaluation analysis are discussed in section 4.

Although a few CLOM have been proposed as reported above, the semantic nature of cross-language mappings that cross-language ontology matching methods are expected to find has not been sufficiently investigated [Abu Helou et al, 2014a]. To provide a semantic interpretation of mappings in CLOM we extended a definition proposed for MOM in a recent work [Atencia et al. 2012].

A *crisp mapping* tells us that a certain concept is related to other concepts in different ontologies and specifies the type of relations, which are typically a set of formal relations $\{\equiv, \sqsubseteq, \text{or} \perp\}$. A *weighted mapping* in addition associates a number (weight) to those relations [Atencia et al. 2012]. Atencia et al. (2012) presented a formal semantics of weighted mapping among independent ontologies, that assumes a classification-based interpretation of mappings.The classification-based approach was considered because many ontology matching methods [Shvaiko and Euzenat, 2007], uses metrics that evaluate the overlap between the entities (e.g., ontology instances, documents, pieces of text) that are "classified" under two concepts. Atencia et al (2012) approach provides a very general definition of classification context (the set of objects considered for the interpretation of mappings), which can support the definition of a formal framework to interpret translations between ontology concepts that are lexicalized in different languages. The classification-based approach can fit the CLOM problem.

In the work carried out in the SIERA project we extended the above definition using the classification-based approach *independently* of the interpretation of classification and the type of objects that can be classified under the concept. To do this we needed to consider the *lexicalization* concept in the classification-based approach, which is a fundamental aspect used by ontology matchers and a central point toward extending such an approach for the CLOM problem.

7.2.2 Preliminaries and Definitions

7.2.2.1 Ontology

Ontologies have gained a lot of attention in recent years as tools for knowledge representation. Ontologies can be defined as a structured knowledge representation system composed of: *classes* (or concepts or topics), *instances* (which are individuals which belong to a class), *relations* (which link classes and instances, allowing to insert information regarding the world represented in the ontology), and *terms* the lexical representation (labels) of the ontology elements in a given natural language.

Definition 1:
*An **ontology**$\mathcal{O}$ is represented as* $\boldsymbol{\mathcal{O} := (\mathcal{C}, \mathcal{R}, \mathcal{T}, \mathcal{I}, \mathcal{E}_R, A)}$*where,*
$\boldsymbol{\mathcal{C}}$is a set of classes (or concepts). $\boldsymbol{\mathcal{R}}$ *is* a set of relations between classes (e.g., hyponymy($\leqslant$), equivalence ($\equiv$), subsumption ($\sqsubseteq$), or disjoint($\perp$)). $\boldsymbol{\mathcal{T}}$is the set of labels of all possible ontology entity (concepts labels, relations labels, comments, among others) in a given natural language ℓ , represented as $\mathcal{T} = \{t_i, \dots t_n\}$ where $t_i \in \ell_j$ such that ℓ_j belongs to a set of languages $\mathcal{L}$, $\ell_j \in \mathcal{L}$. $\boldsymbol{\mathcal{I}}$is a sets of instances where each $i \in \mathcal{I}$ is classified under a class $c \in \mathcal{C}$. $\boldsymbol{\mathcal{E}_R} \subseteq \mathcal{I} \times \mathcal{I}$ is a set of relationships between instances, and $\boldsymbol{A}$ is a set of axioms in a logical language on $\boldsymbol{\mathcal{O}}$.

7.2.2.2 Ontology Matching

With the enormous amount of heterogeneous data in the semantic Web, the field of ***Ontology Matching*** has increasingly become an important research field. Ontology matching, as a solution of semantic heterogeneity, tries to establish correspondences among the semantically related ontological entities.

Definition 2: [Euzenat and Shvaiko 2007]
Ontology Matching is "the process of finding relationships or correspondences between entities of different ontologies"

The ***matching process*** refers to the process of finding relations ($\mathcal{E}_R$ and $\mathcal{R}$) between concepts ($\mathcal{C}$) and instances($\mathcal{I}$) of heterogeneous ontologies. The outcome of this process is referred to as semantic alignment $\mathcal{A}$. The matching process can be viewed as, the set of ontologies to be matched, called ***ontology matching task*** [Euzenat and Shvaiko, 2007], and a set of configurations of a given ontology matching system.

The problem of establishing such relationships consists of operating a certain way of ontology mapping strategies $\mathcal{M}$which that can be either a manual, or a (semi)-automatic, to obtain the alignment result $\mathcal{A}^{\mathcal{M}}$, which is a set of ***correspondences*** between ontological entities [Jung 2007, Euzenat 2008].

In particular, the matching process can be seen as a function $\mathcal{F}$ that takes two (or more) ontologies: the source ontology $\mathcal{O}_S$ and the target ontology $\mathcal{O}_T$ as input[55]. It uses a certain mapping strategy $\mathcal{M}$ to produces a semantic alignment $\mathcal{A}^{\mathcal{M}}$.

This function, the matching process, for a given *ontology matching task* (in our case, consisting of two ontologies) makes use of three matching features, namely: the alignment $\mathcal{A}$, which is to be completed by the process, the matching parameters (it might be empty), $\mathcal{P}$, e.g., simple parameters like weights and thresholds, or complex (e.g., matching task profile as in [Cruz et al. 2012]), and the external resources used by the matching process, r, e.g., common knowledge and domain specific thesauri (See Figure 7.1).
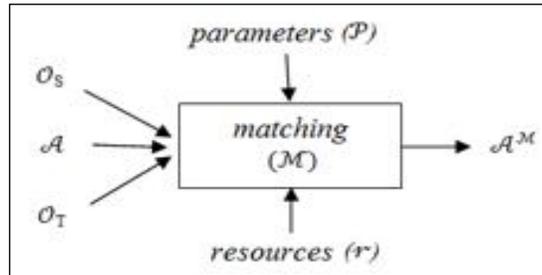
Figure 7.1 : The ontology Matching process [Euzenat and Shvaiko, 2007].

Definition 3: [Euzenat and Shvaiko, 2007]
The **matching process** for a pair of ontologies $\mathcal{O}_S$ and $\mathcal{O}_T$, respectively called source and target ontologies, can be seen as a function $\mathcal{F}$, which takes the two ontologies as inputs, an input alignment $\mathcal{A}$, a set of parameters $\mathcal{P}$, and a set of resources r returns a new alignment $\mathcal{A}^{\mathcal{M}}$ between these ontologies by employing a particular mapping strategy $\mathcal{M}$.

$$\mathcal{A}^{\mathcal{M}} = \mathcal{F}(\mathcal{O}_S \; ; \mathcal{O}_T \; ; \mathcal{A}; \mathcal{P}; r)$$

Definition 4: (Correspondence, also referred to as "mapping"). [Jung et al., 2009].
Given a source ontology $\mathcal{O}_S$, a target ontology $\mathcal{O}_T$, and a set of alignment relations $\mathcal{R}$, a *correspondence* is a quadruple:

$$Correspondence := <c_S; c_T; r; n> c_S \in \mathcal{O}_S, c_T \in \mathcal{O}_T$$

Where r $\mathcal{R}$, a set of alignment relations (e.g., $\equiv$, $\sqsubseteq$, or $\perp$), and $n \in [0, 1]$ is a confidence level (i.e., measure of confidence in the fact that the correspondence holds).

An alignment is a set of mappings expressing the correspondence between two entities of different ontologies through their relation and a trust assessment (confidence value). The relation can be equivalence as well as specialization/generalization or any other kind of relation. The trust assessment can be Boolean as well as given by other measures (e.g., probabilistic or symbolic measures).

---

[55] With the subscript *S* and *T* we refer to the *source ontology* and *target ontology,* respectively.

Definition 5: (Alignment). [Jung et al., 2009]

Once we choose a mapping strategy $\mathcal{M}$ for conducting a matching process, alignment between two ontologies $\mathcal{O}_S$ and $\mathcal{O}_T$ is represented as a set of correspondences;

$$\mathcal{A}_{S,T}^{\mathcal{M}} = \{ <c_S; c_T; r; n> \mid c_S \in \mathcal{O}_S, c_T \in \mathcal{O}_T \}$$

7.2.2.3 Mono, Multi and Cross-Lingual Ontology Matching

A general definition of the ontology matching is proposed in [Euzenat and Shvaiko 2007] (see Definition 1), without explicitly specifying the natural languages used to label the ontology entities. In the literature, the largest part of the ontology matching strategies involve syntactic and lexical comparisons, thus ontologies coming in different languages are very difficult to match.

Ontology entities (e.g., concepts, relations, descriptions, and comments) can be expressed in natural languages, by associating (labeling) them with terms (i.e., a lexicon) that belong to one (or more) natural languages. We denote the notion of **lexicalization** as the process of associating ontology entities with a set of terms that belongs to a set of natural languages, and the notion **lingualization** as the process of retrieving the set of languages that the associated terms belong to.

We can say that an ontology $\mathcal{O}$ is **lexicalized** in a given language ℓ, $\mathcal{T}^{\ell,O}$, if the ontology terms $\mathcal{T}$ are **lexicalized** in language ℓ, such that ℓ belong to the set of natural languages $\mathcal{L}$ ($\ell \in \mathcal{L}$). Ontologies can be lexicalized in one language (so-called, mono-lingual ontology ($|\mathcal{L}| = 1$)), two languages (so-called, bi-lingual ontology ($|\mathcal{L}| = 2$)), or more languages (so-called, multi-lingual ontology ($|\mathcal{L}| > 2$)).

[Spohr et al. 2011] distinguished between the matching tasks based on the number of languages used to lexicalize the ontology terms. Given two ontologies; $\mathcal{O}_S$ and $\mathcal{O}_T$, lexicalized in a set of natural languages $\mathcal{L}_S$ and $\mathcal{L}_T$, respectively, and $\mathcal{T}^{\mathcal{L}_S,O_S}$ and $\mathcal{T}^{\mathcal{L}_T,O_T}$ be the set of terms (labels) of $\mathcal{O}_S$ and $\mathcal{O}_T$ lexicalized in a set of natural languages $\mathcal{L}_S$ and $\mathcal{L}_T$, respectively. Then we can define the following notation:

Definition 6: *Mono-lingual Ontology Matching (MOM),*
MOM is the process of establishing correspondences among ontological resources from two (or more) independent ontologies, where both ontologies are lexicalized in the same natural language. *MOM* is the process of matching entities in $\mathcal{O}_S$ and $\mathcal{O}_T$ by considering the labels in $\mathcal{T}^{\mathcal{L}_S,O_S}$ and $\mathcal{T}^{\mathcal{L}_T,O_T}$ in a *single* language ($\mathcal{L}_S = \mathcal{L}_T$), with $|\mathcal{L}_S \cap \mathcal{L}_T| = 1$.

Definition 7: *Multi-Lingual Ontology Matching (MLOM),*
MLOM is the process of establishing correspondences among ontological resources from two (or more) independent ontologies where each ontology is lexicalized by more than one language; the languages used in each ontology can also overlap.
MLOM is the process of matching entities in $\mathcal{O}_S$ and $\mathcal{O}_T$ by considering the labels in $\mathcal{T}^{\mathcal{L}_S,O_S}$ and $\mathcal{T}^{\mathcal{L}_T,O_T}$ in at least *two* languages, with $|\mathcal{L}_S \cap \mathcal{L}_T| \geq 2$.

Definition 8: *Cross-Lingual Ontology Matching (CLOM),*

CLOMisthe process of establishing correspondences among ontological resources from two (or more) independent ontologies where each ontology is lexicalized in a different natural language(s), one or more natural language, but they do not share any language. *CLOM* is the process of matching the ontological entities in $\mathcal{O}_S$ and $\mathcal{O}_T$ either by *conceptually translating*

a. the labels in $\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S}$to at least one language $\ell' \in \mathcal{L}_T$ and considering the labels in $\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S} \in \ell'$ with those in $\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T} \in \ell'$, or

b. the labels in $\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T}$to at least one language $\ell' \in \mathcal{L}_S$and considering the labels in $\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T} \in \ell'$with those in $\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S} \in \ell'$, or

c. the labels$\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S}$and the labels$\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T}$to at least one language ℓ''such that$(\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S} \cap \mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T}) \in \ell''$ and considering the labels in$\mathcal{T}^{\mathcal{L}_S,\mathcal{O}_S} \in \ell''$ with those in $\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T} \in \ell''$

7.2.2.4 Formal Interpretation of Mappings in Mono-Lingual Ontologies

Atencia et al. (2012) introduced a formal semantics of weighted mapping between different ontologies, based on a classification interpretation of mappings, that is, two concepts are said to be *extensionally equivalent* if the set of objects classified under one concepts can be also (re-)classified under the second concept.

The Atencia et al. approach provides a formal semantics of weighted mapping between logically founded ontologies, which give the notion of logical consequences of weighted mappings that allows to define a set of inference rules to derive a mapping from a set of existing mappings.

"…based on a classification interpretation of mappings: if O1 and O2 are two ontologies used to classify a common set X, then mappings between O1 and O2 are interpreted to encode how elements of X classified in the concepts of O1 are re-classified in the concepts of O2, and weights are interpreted to measure how precise and complete re-classifications are" [Atencia et al. 2012].

Atencia et al. represent a formal semantics for interpreting a confidence value (weight mapping) associated with a mapping. The Atencia et al. approach relies ona classification interpretation of mappings, which takes inspiration from the family of extensional based approaches (for more details on this see [Euzenat and Shvaiko 2007]) used in ontology matching techniques. Atencia et al take advantage of precision, recall, and F-measures, as they are used in the context of classification tasks in their formalization of the weight mapping relation (subsumptions ($\sqsubseteq, \sqsupseteq$) and equivalence ($\equiv$)) that associate mappings to a closed subinterval [a, b], where *a* and *b*are real numbers in the unit interval [0, 1] which respectively define the lower and upper bound that precision and recall fall in .

Intuitively, speaking, suppose we have two ontologies $\mathcal{O}_1$and $\mathcal{O}_2$. Ontology $\mathcal{O}_1$is used to classify the set of elements *{x₁, . . . , x₁₀}*, and suppose the same elements are reclassified in ontology $\mathcal{O}_2$. We can measure the values of the theoretical set of mappings by counting the classified elements. For example, suppose that the elements *{x₁, . . . , x₁₀}*, classified under the concepts $C \in \mathcal{O}_1$ and $D \in \mathcal{O}_2$, then we say that concept C and D are equivalent with a value 0.1

$(\langle C, D, \equiv, 1.0 \rangle)$. Similarly if the elements $\{x_1, \ldots, x_5\}$ classified under the concept $H \in \mathcal{O}_1$ then we have a subsumption relation between H and D with a value 0.5 ($\langle H, D, \sqsubseteq, 0.5 \rangle$).

Staring from the correspondence (i.e., mapping) definition we presented before (see Definition 4), Atencia et al. define the *weighted mapping* as an expression that represents the theoretical set of relation between two concepts belonging to two different ontologies by associating those relations with a closed subinterval of [0,1].

> **Definition 9: Weighted Mapping**
> Given two ontologies $\mathcal{O}_1$ and $_2$, a weighted mapping from $\mathcal{O}_1$ to $\mathcal{O}_2$ is a quadruple:
> $$\text{Weighed Mapping} := <C; D; r; [a,b] >$$
> where $C \in \mathcal{O}_1$ and $D \in \mathcal{O}_2$, $r \in \{\sqsubseteq, \equiv, \sqsupseteq, \perp\}$, and a, b are real numbers in the unit interval [0, 1].

Following the standard model-theoretic semantics based on interpreting classes as sets: an interpretation $\mathfrak{T}$ is a pair $\mathfrak{T} = \langle \Delta^{\mathfrak{T}}, \cdot^{\mathfrak{T}} \rangle$ where $\Delta^{\mathfrak{T}}$ is a non-empty set, called domain of interpretation $\mathfrak{T}$, and $\cdot^{\mathfrak{T}}$ is a function that interprets each concept (class) $C \in \mathcal{C}$ as a non empty subset of $\Delta^{\mathfrak{T}}$, and each instance identifier ($x \in X$) as an element of $\Delta^{\mathfrak{T}}$.

> Given an ontology $\mathcal{O}$, let $\mathcal{C}$ be a set of concepts, $\mathcal{R}$ a set of relations, and X a set of shared objects.
> Then $C^{\mathfrak{T}} \in \Delta^{\mathfrak{T}}$ for $C \in \mathcal{C}$, $r^{\mathfrak{T}} \in \Delta^{\mathfrak{T}} \times \Delta^{\mathfrak{T}}$ for $r \in \mathcal{R}$, and $x \in \Delta^{\mathfrak{T}}$ for $x \in X$.

Suppose that the concepts of two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ are used to classify a common set of elements X. Then the mappings between concepts in $\mathcal{O}_1$ and $\mathcal{O}_2$ encode how the elements of X classified under the concepts of $\mathcal{O}_1$ are re-classified in the concepts of $\mathcal{O}_2$, and the weights encode how precise and complete these re-classifications are.

> [Atencia et al. 2012]: *"Let $X = \{x_1, \ldots, x_n\}$ be a non-empty finite set of fresh constants not occurring in $L(O_1)$ or $L(O_2)$. The set X is meant to represent the set of shared items classified by concepts of the ontologies O_1 and O_2. A classification of X in O_1 is specified by virtue of an interpretation $\mathfrak{T}_1$ of O_1 extended with the elements of X as follows."*

Let C be a concept of O_1 and x_k a fresh constant of X; we define X as a *shared context* (domain) of the mapping. We say that x_k is classified under C according to $\mathfrak{T}_1$ if $x_k^{\mathfrak{T}_1} \in C^{\mathfrak{T}_1}$. Then, the set $C_X^{\mathfrak{T}_1} = \{x \in X | x^{\mathfrak{T}_1} \in C^{\mathfrak{T}_1}\}$ represents the subset of items of X classified under C according to $\mathfrak{T}_1$.

Note that $C_X^{\mathfrak{T}_1}$ is a subset of X ($C_X^{\mathfrak{T}_1} \subseteq X$), whereas $C^{\mathfrak{T}_1}$ is a subset of the domain of the interpretation $\mathfrak{T}_1$ ($C^{\mathfrak{T}_1} \subseteq \Delta^{\mathfrak{T}_1}$). In addition, $C_X^{\mathfrak{T}_1}$ is always a finite set, while $C^{\mathfrak{T}_1}$ may be infinite.

Let $\mathfrak{T}_1$ and $\mathfrak{T}_2$ be interpretations of O_1 and O_2, respectively, and let C and D be the concepts of O_1 and O_2, occurring in the correspondence $\langle C, D, r, [0,1] \rangle$. The sets $C_X^{\mathfrak{T}_1}$ and $D_X^{\mathfrak{T}_2}$ can be compared as they are both subsets of X which represents the sets of items of X classified under C according to $\mathfrak{T}_1$ and under D according to $\mathfrak{T}_2$, respectively. Then the different types of mappings $\langle C, D, r, [0,1] \rangle$ obtained by looking at the different $r \in \{\sqsubseteq, \sqsupseteq, \equiv, \perp\}$.

Intuitively, following the classification tasks, the mapping $\langle C, D, \sqsubseteq, [0,1]\rangle$ is used to express that any item in X which is classified under C according to $\mathfrak{T}_1$ is (re-)classified under D according to $\mathfrak{T}_2$. The confidence level interval $[0,1]$ (the weighted mapping, [Atencia et al. 2012]) can be seen as the recall of $C_X^{\mathfrak{T}_1}$ w.r.t $D_X^{\mathfrak{T}_2}$.

$$R\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = \frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|C_X^{\mathfrak{T}_1}\right|} \in [a, b]$$

In the same way, the mapping $\langle C, D, \sqsupseteq, [0,1]\rangle$ is used to express the fact that the fraction of items of X classified by D according to $\mathfrak{T}_2$ which are (re-) classified under C according to $\mathfrak{T}_1$. The confidence level (weighted mapping) can be seen as the precision of $D_X^{\mathfrak{T}_2}$ w.r.t $C_X^{\mathfrak{T}_1}$.

$$P\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = \frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|D_X^{\mathfrak{T}_2}\right|} \in [a, b]$$

By keeping parallelism with classification systems, the natural way to interpret the confidence level (weighted mapping) of the equivalent relation that aligns two concepts C and D, $\langle C, D, \equiv, [0,1]\rangle$, is by means of the F-measure, which is the harmonic mean of precision and recall. Typically the F-measure used to evaluate the global quality of a classifier, the *F-measure* of $C_X^{\mathfrak{T}_1}$ and $D_X^{\mathfrak{T}_2}$ is defined as

$$F\left(C_X^{\mathfrak{T}_1}, D_X^{\mathfrak{T}_2}\right) = 2.\frac{\left|C_X^{\mathfrak{T}_1} \cap D_X^{\mathfrak{T}_2}\right|}{\left|C_X^{\mathfrak{T}_1}\right| + \left|D_X^{\mathfrak{T}_2}\right|} \in [a, b]$$

An interesting point in the Atencia et al. weighted mapping definition is the use of ranges of scores [a, b] for subsumption relations that are interpreted as the precision $\langle C, D, \sqsubseteq, [a, b]\rangle$, and recall $\langle C, D, \sqsupseteq, [a, b]\rangle$. By this we can define the equivalence relation as a conjunction of the two subsumption relations. This in particular gives the notion of logical consequences of weighted mappings that allows defining a set of inference rules to derive a mapping from a set of existing mappings.

For instance, if we have weighted mappings $\langle C, D, \sqsubseteq, [a, b]\rangle$ and $\langle C, D, \sqsupseteq, [e, f]\rangle$, then we can derive the equivalence weighted mapping $\langle C, D, \equiv, [v, w]\rangle$ with $v = \min(a, e)$ and $w = \max(b, f)$. Notice that, if we consider the usual definition of equivalence in DLs in terms of subsumption: $\langle C \equiv D\rangle$ iff $\langle C \sqsubseteq D\rangle$ and $\langle C \sqsupseteq D\rangle$, when dealing with single values for precision($\sqsubseteq$) and recall($\sqsupseteq$) instead of intervals, it is usually impossible to combine them into a single value by simple conjunction [Atencia et al. 2012].

Nevertheless, generally ontology matchers are used to return a single confidence level value, for instance, *n*. Accordingly, to represent the value *n* by means of the weighted mapping interval [a, b], the authors [Atencia et al. 2012] suggest to use apointwise interval; we can assume that *a=b*, then *n=[a, a]*. Thus, we can simply present the mapping relation as $\langle C, D, r, n\rangle$.

Figure 7.2, demonstrates the extensional meaning between two concepts *C* and *D* of the ontology $\mathcal{O}_1$ and ontology $\mathcal{O}_2$ respectively, based on the classification-based mapping approach. $\mathfrak{T}_1$ and $\mathfrak{T}_2$ represent an interpretation of $\mathcal{O}_1$, and $\mathcal{O}_2$, respectively. $\Delta^{\mathfrak{T}_1}$ and $\Delta^{\mathfrak{T}_2}$ represent the domain of interpretation of $\mathfrak{T}_1$ and $\mathfrak{T}_2$, respectively. The set $C_X^{\mathfrak{T}_1}$ and $D_X^{\mathfrak{T}_2}$ represent the subsets of items of X classified under C according to $\mathfrak{T}_1$, and under D according to $\mathfrak{T}_2$, respectively. Objects *z* and *y* represent an objects do not belong to the shared domain X.
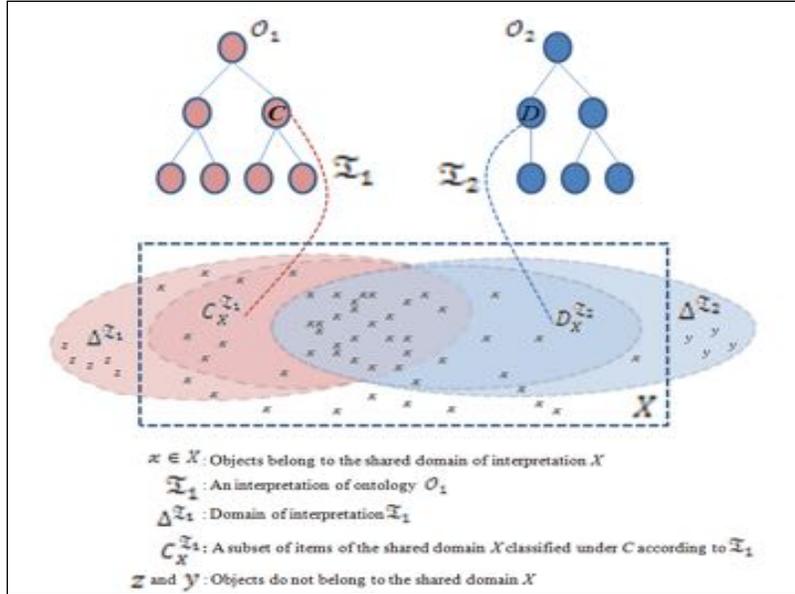
Figure 7.2: The extensional meaning of a concepts [Abu Helou et al. 2014a]

7.3 Framework for Mapping Between WordNet and Arabic Ontology

7.3.1 Mapping concepts across different languages

In the context of cross-lingual mapping, languages barrier has been attempted by transforming a cross-lingual mapping problem into a mono-lingual mapping one by leveraging translation tools [Spohr et al.,2011, Fuet al. 2012]. However, the cultural-linguistic barriers [Gracia et al. 2012] still need more efforts in terms of the mapping process and techniques, as well as to formally define the mappings semantic that align concepts lexicalized in different natural languages.

Mapping concepts that are lexicalized in different languages is a challenging task, without loss of generality, if two concepts are lexicalized in different languages, then they are considered equivalent if they express the same meaning in a given context (i.e., same concept). If two language communities (the majority of language speakers) share the same understanding for a given concept, whatever the lexical notation being used (language), as they refer to the instances that are classified under this concept.

The goal is to provide a formal definition of the cross-language ontology matching problem, mainly to define what a correspondence is, and how to represent correspondences in cross-language ontology mapping (CLOM) problem, that is, to define the semantics of the correspondence considering lexicalization in the definition of the mapping.
To achieve this, we need to define a formal interpretation (i.e., formal semantics) of the correspondence in the CLOM. For this, we extend the formal definitions of the classification-based semantic, which we found suitable for our purposes.

The classification-based interpretation fits our problem because many of the approaches working in CLOM were founded on the extensional approach often based on statistics because most of the machine translation tools are statistically based. Since the classification-based approach defines semantic based on the use of the concepts and the classes for objects, we believe that this approach is useful in our case because it can also support the definition of translation between concepts in ontologies that are lexicalized in different languages, based on the use of this concept as classifiers. Also many ontology matching methods are experimented on datasets of information objects (e.g., documents, video, images, etc) that are classified with under concepts.

At the same time, the adaptation of a CLOM framework is not trivial. The Atacia et al. approach does not explicitly introduce the lexicalization of the ontologies, which is important for it to be used by ontology matchers [Euzenat and Shvaiko 2007]. As a result, we provided a lexicalized version of the ontology matching problem mainly in the context of CLOM based on classification interpretation of mappings.

Next, we present the foundation of the proposed cross-language ontology matching framework. Practically, we define what is the semantics of a correspondence for cross-language ontology matching, taking into account the lexicalization in the definition of the mapping. We define the notion *semantonym* ; a cross-langauge mapping based on a classification-based approach. We provide an overview of an experimental setting, through which the proposed mapping semantics can be evaluated and a gold standard dataset can be generated. We also demonstrate the usage of the proposed framework by mapping 10,000 Arabic concepts dataset in the Arabic Ontology to their equivalent concepts in WordNet. Such datasets can be used to as reference alignments to assess the quality and to compare alternative cross-language mapping methods.

7.3.2 Classification-based Interpretation for Cross-Lingual Mappings

The extension of concept is often used in many cross-language ontology matching strategies [Euzenat and Shvaiko, 2007]; this extension is interpreted in different ways, e.g., instances classified under concepts, or even a document annotated with a concept. We believe this is a promising approach to provide a foundation to CLOM, and it makes sense to adopt such an approach that is based on the classification of different kind of objects with a concept to interpret the semantics of mapping.

We interpreted the classification task as a task to establish whether an instance i is member of a class C, i.e., if i belongs to the extension of C. This extensional interpretation cannot be directly applied for ontologies that are not formally (logically) represented and interpreted in set theoretic semantics. If we consider a sentence and we want to disambiguate the meaning of the words in it, we can consider the disambiguation task as a form of classification, namely, the classification of a word as occurrence of a word sense in the sentence.

We extended the notion of a mono-language matching definition to the cross-language matching one by considering the *lexicalization* of the ontology entities; we adapted the approach presented in [Atencia et al.2012] that provides a formal interpretation of the semantics for the weighted ontology mapping based on the extension of concepts. Let $X = \{x_1, ..., x_n\}$ be

a non empty finite set of instance constants, and let C_ℓ be a concept lexicalized in language ℓ; we say that instance x_n is classified under C_ℓ according to $\mathfrak{T}_1$ if $x_n^{\mathfrak{T}_1} \in C_\ell^{\mathfrak{T}_1}$. Then, the set $C_\ell^{\mathfrak{T}_1} = \{x \in X \mid x^{\mathfrak{T}_1} \in C_\ell^{\mathfrak{T}_1}\}$ represents the subset of instances belonging to X classified under the lexicalized concept in a given language ℓ, C_l, according to the interpretation $\mathfrak{T}_1$. Note that $C_{X,\ell}^{\mathfrak{T}_1} \subseteq X$ and $C_\ell^{\mathfrak{T}_1} \subseteq \Delta^{\mathfrak{T}_1}$.

Let $\mathfrak{T}_1$ be interpretation of ontology $\mathcal{O}_1$ lexicalized in language ℓ, and $\mathfrak{T}_2$ be interpretation of ontology $\mathcal{O}_2$ lexicalized in language ℓ'. And let C and D be lexicalized concepts of $\mathcal{O}_1$, and $\mathcal{O}_2$, respectively, occurring in the correspondence mapping $\langle C_\ell, D_{\ell'}, r, [a, b] \rangle$.

Then, the sets $C_{X,\ell}^{\mathfrak{T}_1}$ and $D_{X,\ell'}^{\mathfrak{T}_2}$, can be compared as they are both subsets of X, which represent the sets of objects of X classified under the lexicalized concept C_ℓ according to $\mathfrak{T}_1$ and under the lexicalized concept $D_{\ell'}$ according to $\mathfrak{T}_2$, respectively. Following the classification-based mapping proposed in [Atencia et al.2012], we interpret the confidence level of the extensional equivalent relation by means of the F-measure, which is the harmonic mean of precision and recall.

$$\text{The } \textit{F-measure} \text{ of } C_{J,\ell}^{\mathfrak{T}_1} \text{ and } D_{J,\ell'}^{\mathfrak{T}_2} \text{ is defined as } F\left(C_{X,\ell}^{\mathfrak{T}_1}, D_{X,\ell'}^{\mathfrak{T}_2}\right) = 2 . \frac{\left|C_{X,\ell}^{\mathfrak{T}_1} \cap D_{X,\ell'}^{\mathfrak{T}_2}\right|}{\left|C_{X,\ell}^{\mathfrak{T}_1}\right| + \left|D_{X,\ell'}^{\mathfrak{T}_2}\right|}$$

As discussed before, the weight mapping between two objects is by means of an interval [a, b], while in general the ontology matching algorithm used to return a single confidence level value, for instance, *n*. Accordingly, to represent this value *n* by means of the weighted mapping interval [a, b], a point-wise interval can be used; that is, we assume that *a=b*, then *n=*[a, a]. Thus, we can simply present the mapping relation as $\langle C_\ell, D_{\ell'}, r, n \rangle$.

7.3.3 The Semantonym Mapping

In this section, we introduce our notion for semantic mapping between two concepts lexicalized in different languages, called ***semantonym***, a cross-language mapping based on a classification-based approach.

Intuitively, two concepts lexicalized in different languages are considered to be *semantonym* (i.e., same concept). If a community of language speakers agrees that the extension of both concepts are correctly applied in a given context, then we can say that the extension of the concept C_s and the extension of the concept C_T are equivalent with a confidence level *n*. Various approaches can be adopted to measure the confidence level based on the interpretation of the extensional model. For instance, a probabilistic one can be adapted similar to [Atencia et al. 2012] as introduced above for the well-founded logical ontologies.

We hypothesize that in order to share a meaning (concept) we have to share a domain of interpretation, and this domain represents the shared context of a community of languages speakers. Considering the extensional based approach, particularly the case of cross-lingual extensional meaning of a concept, we should keep in mind that according to a given shared context, it is *not* necessary that all objects classified under C_S ($x \in C_{X,S}^{\mathfrak{T}_1}$) are also instances under D_T ($x \in D_{X,T}^{\mathfrak{T}_2}$) according to an interpretation $\mathfrak{T}_1$ and $\mathfrak{T}_2$, respectively. It happens that an

object$x \in C_{X,S}^{\mathfrak{T}_1}$ might *not* exist in the other language (or, ontology) $(x \notin D_{X,T}^{\mathfrak{T}_2})$, or even it might be classified under another concept such as $(x \in E_{X,T}^{\mathfrak{T}_2})$.

Definition 9: *Cross-lingual correspondence (Semantonym)*
Given two lexicalized concepts $c_S \in \mathcal{O}s$, and $c_T \in \mathcal{O}_T$ in $\mathcal{O}_S$ and $\mathcal{O}_T$ respectively, and lexicalized in ℓ_S and ℓ_T the language respectively.
Then c_S is a *semantonym* of $c_T (c_S, \mathcal{S}, c_T)$ with a confidence $n \in [0,1]$, $\langle c_S, c_T, \mathcal{S}, n \rangle$ if they are extensionally equivalent (in a given context), using a certain mapping strategy $\mathcal{M}$.

Definition 10: *Cross-lingual alignment*
Given two ontologies, $\mathcal{O}_S$ and $\mathcal{O}_T$, lingulized in language ℓ_S and ℓ_T, respectively, a conceptual cross-lingual alignment through the conceptual translation and a particular mapping strategy $\mathcal{M}$ is a set of correspondences: $\mathcal{A}^{\mathcal{M}} = \{<c_S; c_T; \mathcal{S}; n> \mid c_S \in \mathcal{O}_S, c_T \in \mathcal{O}_T\}$ where $\mathcal{S}$ is the semantonym relation, and n the confidence level for each correspondence pair. $\in$

The extensional equivalence between two concepts represents the common shared knowledge between the community of language users (speakers), so-called shared context (domain) of interpretation. The confidence level of the mapping relation (the semantonym) can reveal the acceptance level of the equivalent relation based on a given threshold (e.g., $n \geq 0.95$).
Based on the classification-based interpretation of mapping in a logical domain, two concepts lexicalized in different languages are said to be *semantonym* if most of the objects classified under the first concept can also be classified under the second one in a given context.

Proposition 1.
Given two lexicalized concepts C and D in $\mathcal{O}_S$ and $\mathcal{O}_T$, respectively, and $\mathcal{L}_S$ and $\mathcal{L}_T$ the set of associated languages in $\mathcal{O}_S$ and $\mathcal{O}_T$, respectively, then c_s is *semantonym* of c_T : $(c_s \mathcal{S} c_t)$ if the F-measure of the extension of c_s and the extension of c_T is greater than a certain threshold.

$$F\left(C_{X,\ell}^{\mathfrak{T}_1}, D_{X,\ell'}^{\mathfrak{T}_2}\right) = 2 \cdot \frac{\left|C_{X,\ell}^{\mathfrak{T}_1} \cap D_{X,\ell'}^{\mathfrak{T}_2}\right|}{\left|C_{X,\ell}^{\mathfrak{T}_1}\right| + \left|D_{X,\ell'}^{\mathfrak{T}_2}\right|} > THRESHOLD$$

Where $\ell \in \mathcal{L}_S$ and $\ell' \in \mathcal{L}_T$, and X is the set of object in the shared domain.

Following the *classification-based* approach, we can define our notion of *semantonym* using the classification-based approach which can be ground on different ways to characterizing the classification problem.
In view of this, concepts can be associated with extensions (instances representation) in different ways depending on the type of ontology they represent. Formally, we present the *extension of a concept*, recalling that the set $C_{X,L}^{\mathfrak{T}_1} = \{x \in X \mid x^{\mathfrak{T}_1} \in C_L^{\mathfrak{T}_1}\}$ represents the subset of objects belonging to a shared context Xc classified under the lexicalized concept C_L in a given set of languages $\mathcal{L}$, according to the interpretation $\mathfrak{T}_1$. Based on this, several approaches of classification-based interpretation can be adopted to identify the *extension of a concept*, for instance;

- *logical-based:* it is ontology instances (objects), where each one can be classified under the concept *c* (i.e., a named entity classified as *C*).
- *Corpus-based*: it is a corpus of documents, where each term in a given documentcan be classified (annotated)with the concept *c*, *sense(t)*=C, that is, the terms that convey the concept, i.e., the intended meaning of the term *t* in a document.

7.3.4 Experiment design on cross-language mapping validation

In this section we present an experimental setting that is aimed at showing that the above described semantics can be used, in principle, to define cross-language ontology alignments by assigning classification tasks to bilingual speakers. This is important because a gold standard is needed to comparatively evaluate alternative cross-language mapping methods and few high-quality gold standards are available at present [Abu Helou et al. 2014a].

Consider a corpus of sentences, where each sentence expresses a context and a word in the sentence represent the usage of a concept. If a majority of speakers (i.e., bilingual native speakers or lexicographers) can substitute two words, each belonging to a different language, in a sentence and both words indicate the same sense (meaning), then they can be used interchangeably to refer to the same concept (word sense).

We hypothesize that, if speakers can substitute two words in a given context, then these words are synonyms and give an equivalent meaning (concept) [Miller and Fellbaum 1991]. This is valid also for intra- and inter-lingual substitution, as concepts are independent of specific languages. We assume the above hypothesis but, instead of considering the cross-language substitutability of words themselves, we consider the cross-language substitutability of meanings associated with these words, by referring to *co-disambiguation* (see definition 11, [Abu Helou et al. 2014a]) of words across ontologies in different languages.

> **Definition 11**: *Co-disambiguation task,* let *WSD(w_i)* be a function called Word Sense Disambiguation, such that w_i is an occurrence of the word *w* in a sentence *S*. *WSD* associates w_i with a sense in a lexicon (e.g., WordNet). Accordingly, we can define a *cross-language WSD* function *CL-WSD$_{[L1>L2]}$(w_i)*, such that *CL-WSD* associates a word w_i in a language L_1 (where $L_1$is the language used in *S*) with a sense in a lexicon lexicalized in another language L_2.

In another words, if the substitution of the words does not change the meaning of the context, then they are conceptually equivalent. In view of this, *CL-WSD* can be seen as a classifier, where the number of agreements among the lexicographers (bilingual speakers) expresses the confidence (i.e., the weight) of the mapping.

The speakers perform the *CL-WSD* tasks, and the mapping between two word senses depends on a frequency-based function that measures the degree in which the two senses in two different languages co-disambiguate the same word sense in multiple contexts (sentences). Suppose we have a corpus of English sentences, we find a word w_{en} that appears in these sentences. We

disambiguate each occurrence of $w_{en,i}$ with an English word sense C_i; we disambiguate each occurrence of $w_{en,i}$ with a synset D_i in Arabic. As a result of this operation we found two sets of distinct concepts $\bar{C}$ and $\overline{D}$ that have been used to disambiguate w_{en} respectively in English and Arabic. For each $C_i \in \bar{C}$ we count the number of $_i$ that has been co-disambiguated with every $D_i \in \overline{D}$. The co-disambiguation fraction of the two concepts C and D represent the degree at which we can consider C *as a subclass of* D.

Although we use a classification task that differs from the one proposed in (Atencia et al. 2012), we can still use the inference rule they proposed to reason about mappings, to infer new mappings from existing mappings. Moreover, using the *CL-WSD* function as a classification task to evaluate the existence of relations among concepts, we can define a method to establish reference relationships between concepts by performing *CL-WSD* on sentence corpuses.

Proposed Experiment: in order to validate the equivalent relation we need to perform the following *CL-WSD* classification tasks: given a parallel corpus (or two corpuses) which lexicalized in English and Arabic. We disambiguate each occurrence of $w_{en,i}$ in English sentences with a word sense C_i and D_i in English and Arabic respectively. In this way, we obtain two sets of distinct concepts $\bar{C}$ and $\overline{D}$ that have been used to disambiguate the English word w_{en} respectively in senses form English and Arabic. For each $C_i \in \bar{C}$ we count how many times C_i has been co-disambiguated with every $D_i \in \overline{D}$. The co-disambiguation count for the two concepts C and D represent the degree (confidence level) at which we can consider C *as a subclass of* D.

In the same way, we disambiguate each occurrence of $w_{ar,i}$ in Arabic sentences with a word sense C_i and D_i in English and Arabic respectively. The distinct set of concepts $\bar{C}$ and $\overline{D}$ have been used to disambiguate the Arabic word w_{ar} respectively in senses from English and Arabic. For each $D_i \in \overline{D}$ we count the number that D_i has been co-disambiguated with every $C_i \in \bar{C}$. The proportion of the co-disambiguation for the two concepts D and C represent the confidence level at which we can consider D *as a subclass of* C. Then we use the F-measure to interpret the confidence level of the equivalent relation that aligns the two concepts C and D.

However, it might be difficult and costly to make such experiment at large scale. One way is to use available sense annotated corpuses. Nevertheless, such an Arabic corpus is not available. Therefore, we propose to mine the subclass relations starting form a sense annotated English corpus, we *CL-WSD* the English words with the equivalent Arabic senses, and then we check if these relations can be converted to equivalence relations by exploiting the structure (relations) of the WordNet.

The proposed experiment corresponds to a classification task; asking bilingual speakers to perform a *CL-WSD$_{[En>Ar]}$* classification task. We collect sentences from *"Princeton Annotated Gloss Corpus"*, a corpus of manually annotated WordNet synset definitions (glosses). The selected sentences are annotated with at least one sense that belongs to *"Core WordNet"*. The reason for selecting Core WordNet concepts is that they represent the most frequent and salient concepts and thus can shared among many or most languages. Accordingly, we hypothesize that mapping the core WordNet concepts to the equivalent Arabic concepts will form the core for the Arabic Ontology. Then we can extend it to include more cultural and language-specific concepts.

> **The act of starting to construct a** *home₂ⁿ*(BAET, بيت)
>
> *hypernym*↑   ?↑
>
> **The act of starting to construct a** *house₁ⁿ*(MNZL, منزل)
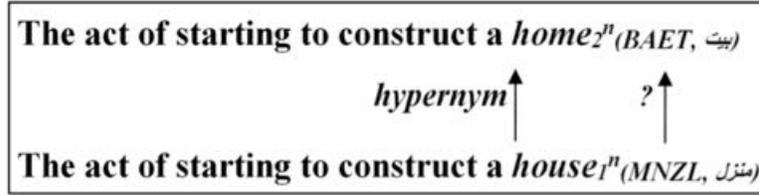
Figure 7.3: Example of *CL-WSD* task and a possible inference.

For each English word sense, a number of bilingual speakers (lexicographers) are asked to provide the equivalent Arabic word sense. For each word sense, the lexicographers substitute the English word with one of the Arabic synsets, which have been developed at Sina Institute and classified under the top levels [Jarrar et al, 2013]. Using available bilingual dictionaries the lexicographers select the best translation. In Figure 4, in the sentence "the act of starting to construct a *house*", the English word "house" was *CL-WSD* with the English sense $house_1^n$ and the Arabic sense (منزل, Mnzel)[56]. For the same sentence we substitute the sense $house_1^n$ with its direct hypernym (subclass) sense $home_1^n$ from the WordNet. We *CL-WSD* the sense $home_1^n$ with the Arabic sense (بيت, Baet).Ideally, we should be able to deduce the subclass relation between (منزل) and (بيت).

However, as mentioned before, not every concept is lexicalized in both (all) languages. The mappings thus obtained will form an initial semantic network. However, conflicts and overlaps might exist. The top levels concepts can [Jarrar et al, 2013] control and eliminate part of this problem. For example, the associated concepts should be classified under the same top concept. This direction of work also taking into account the relations confidence level will be pursued in the future.

Figure 7.4 graphically depicts the experiment proposed to to define cross-language ontology alignments, and in particular reference alignments, by assigning classification tasks to bilingual speakers. We propose to consider the Core WordNet concepts (5,000 concepts), which should represent shared concepts among different languages [Graber et al. 2006]. The majority of speakers is simulated by incorporation larger number of bilingual speakers (lexicographers). We suggest adopting a crowdsourcing method (e.g., Amazon Mechanical Turkey (Sarasua et al. 2012) to collect feedback from larger number of lexicographers.

A significance result of a full-scale version of the proposed experiment is to generate a gold standard for cross-language mappings, which can be used to assess the various automatic cross-language matching systems as well to validate the proposed semantic mapping framework.

---

[56] Translation was obtained using Wikipedia inter-lingual links.
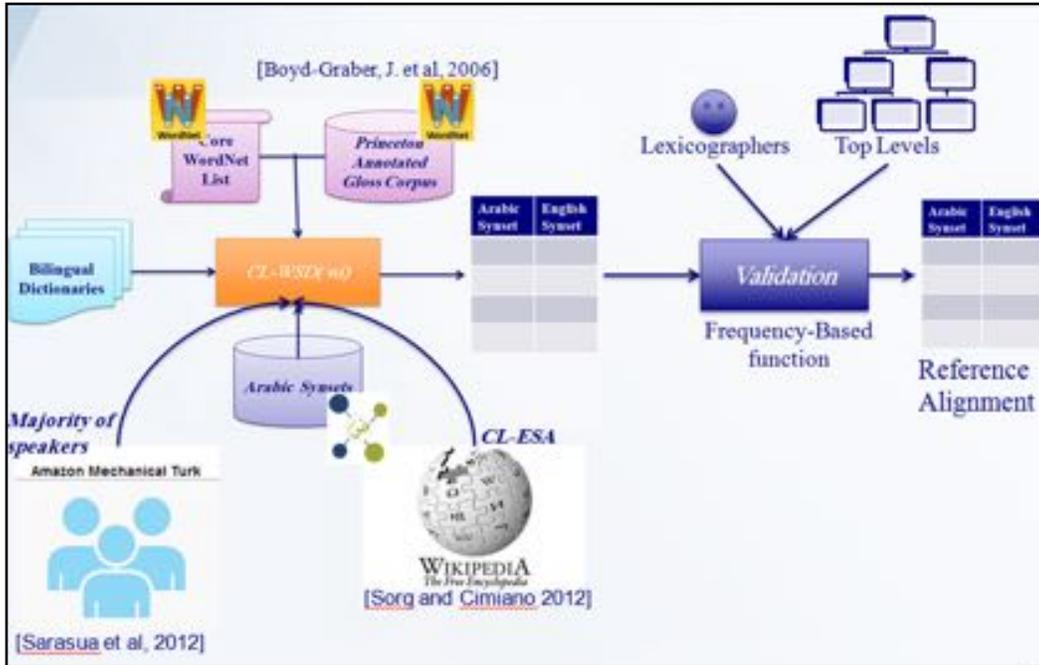
Figure 7.4: using the CLOM framework for building reference alignment

7.4 Cross-Language Mapping Algorithm

The manual construction of ontologies, and in particular ontologies covers natural languages, is expensive and time-consuming task. Automatic construction of wordnets is another method for building and linking wordnets.

To create large linguistic ontologies via cross-language matching approach one needs to map an unstructured or a weakly structured lexicon, to a structured lexicon [Jarrar 2011, 2012], this introduce an extremely difficult matching problem, to mention few of these reasons:

- the lack of structural information,
- the ambiguity of words due to the polysemy and synonymy of word,
- the quality and coverage of the translation sources, and
- the large mapping space (e.g., WordNet has 117659 concepts).

To solve these problems we need a semi-automatic approach that support users (e.g., expert or crowd as workers) to select the most appropriate mappings among relevant candidate mappings. However, before one can select and/or extend the more appropriate existing cross-language ontology matching techniques, we need to be able to compare alternative methods and to assess the quality of their output. We proposed a semi-automatic mapping framework that tries to map concepts (synsets) in different languages by combining translation tools and word sense disambiguation (WSD) into a hybrid task. Inspired form crowdsourcing framework presented in [Venetis et al 2012] we define a mapping algorithm for constructing linguistic ontologies,

through mapping unstructured concepts (i.e., has no relations among concepts) to structured one, as a maximization problem that retrieve *top-k* mappings from a set of sorted candidate mappings.

The mapping algorithm finds the equivalent mappings by ranking the translated synsets, this is performed based on the frequency of translated synsets and majority voting approaches. The algorithm can be viewed into two folds; translation and sense selection (disambiguation) tasks.

- The translation task have to deal mainly with two constrains, (i) the coverage of the translation: assuming that all sense distinctions given by the translations are available in a translation source (e.g., machine translation (MT), machine readable dictionaries (MRD), and multilingual/parallel corpus). (ii) the correctness of the translation: the translation is considered to be correct if it preserves the meaning of the word in context in the source language. We adopt from the CLOM field (e.g., Fu et al. 2012) the translation-based approach in order to overcome the linguistic gaps. Bilingual Machine Readable Dictionaries (MRD) and Machine Translation (MT) tools were used to translate the source synsets into the target synsets.

- The disambiguation task, is to identify the intended meanings of words (word senses) in context, that is, to select the most commonly and accepted meaning of a word [Navigli 2009]. The difficulties of this task rise due to the words polysemy (if it can convey more meanings) and synonymy (if it can convey the same meanings). To resolve this issue we borrowed from the information retrieval filed the frequency based counting (bag-of-word) approach, we also applied a weighted voting to consider the importance of the translation.

The algorithm leverage on translation tools and tries to map synsets lexicalized in one language (e.g., Arabic) to their correspondence synsets in other language (e.g., English), different translating settings were considered to investigate the appropriate translation methods for obtaining the correct translation.We also ranked the translated synsets in order to select the most appropriatesenses. We benchmarked our algorithm on two standard datasets, namely; the Arabic WordNet [Rodrguez et al. 2008] and the Italian component of the MultiWordnet [Piantaet al. 2002], this is followed by a deep experimental analysis and discussion. By this experiment we aim to respond on the following questions:

- what is the best translation tools in term of coverage and correctness.
- what is the impact of the correct translation on the sense disambiguation task.
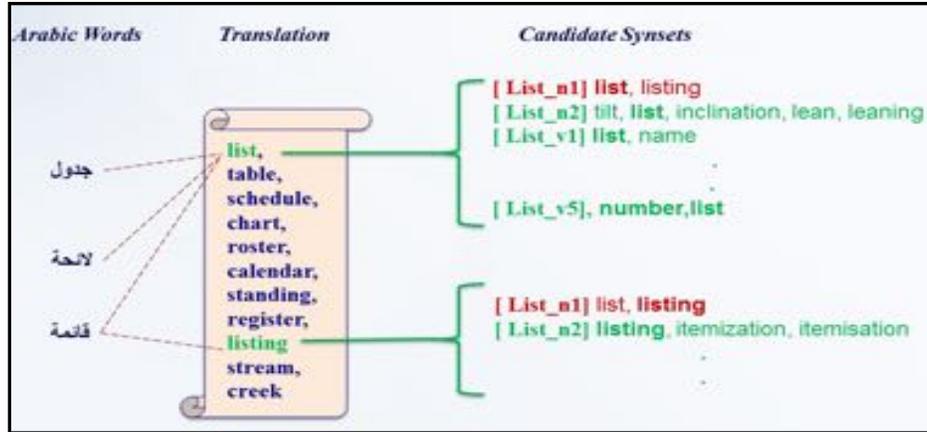
Figure 7.5: example of mapping Arabic sysnets to WordNet

Figure 7.6 illustrates the mapping algorithm by collecting candidate mappings for the Arabic synset {جدول ، لائحة ، قائمة}. All possible translations in English for each Arabic word are collected (e.g., using Google translator). Then we search (lookup) the possible senses from WordNet (candidate senses) for each translated word. After that, we rank the candidate senses, the ranking is performed by counting the frequency of each sense For instance, the sense ($list_1^n$) is obtained by the synonym translation (list and listing), a higher priority (weight) is given to the candidate senses which obtained from translating different words in the source synset, the three words were translated as (*list*) , see the red dashed lines). At the end, selecting the *top-k* mappings (a set of mappings), the appropriate mapping should be among them. The minimum value of *k,* such that it provides the equivalent mappings, the better performance is obtained by the algorithm. our objective is to maximize the number of equivalent mappings by reducing the value of *k*.

7.4.1 Experimental Evaluation

We carried out several experiments to evaluate the effectiveness of the proposed mapping algorithm and, in particular, to evaluate (i) the coverage of the machine translation tools and dictionaries for the identification of candidate matches; (ii) the limits of state-of-the-art ranking methods based on voting strategies.

Three wordnets have been used in the experiment; the Arabic WordNet (2.0) [Rodrguez et al. 2008], the Italian component of the MultiWordNet [Pianta et al. 2002], and the English WordNet (3.0); the three wordnets, respectively, have 11214, 25584 and 117659sysnets. 15964, 40178 and 155287 words. 23481, 41494 and 206941 total word senses. We benchmarked our algorithm with the Arabic and Italian wordnets because large sets of mappings exist from the Arabic WordNet (2.0) and the Italian component of the MultiWordNet to the English WordNet . Thus, we can evaluate the performance of the mapping algorithm at large scale using well-known measures defined in Information Retrieval such as Precision, Recall and F-Measure [Navigli 2009].

It should be noticed that the quality of these wordnets is uncertain; in particular, the Arabic WordNet contains concepts that have been obtained by translating the most frequent English

WordNet concepts in Arabic. This can introduce a bias in our results. However, the results are still valuable in defining upper and lower bounds for coverage and accuracy in a best case scenario.

We evaluated the disambiguation task using evaluation measures borrowed from the information retrieval field [Navigli 2009].
In the experiment different source for *translation* were used;

(i) Machine Translation (MT) tools: we obtained Google translations for all the Arabic (ArWN)and Italian (ItWN) words.

(ii) Machine readable dictionaries (MRD): we used Sina dictionary which is a result of the ongoing Arabic Ontology project, the dictionary was constructed by integrating several specialized and general domain dictionaries.

(iii) Oracle translation (correct translation): An oracle translation was used to demonstrate theupper bounds and to specify the highest expected performance of the proposed approach. An oracle is a hypothetic system which is always supposed to know the correct answer (i.e., the correct translation). We used the translations provided in the benchmark wordnets as an oracle (correct translation). Moreover, an *extend* oracle translation was obtained for the Arabic-English translations by considering all the synonyms of the translated word, not only translation provided in the ArWN.

We performed our experiments considering the two benchmark datasets the ArWN and ItWN. The results are reported in Figure 4 (top); the "Experiment" column specifies the translation method, we also report (in %) the upper bound (max value at k=100) of the evaluation measures: coverage, precision, recall and F-measure. The lower bounds (baseline First sense heuristic [Navigli 2009]) experiments were also reported. Figure 7 (bottom) compares the precision of the different translation methods; the reported measures evaluate if the equivalent mappings are among the top-k ranked mappings (k $\in$[1; 100]). Four variants that exploit the *structural information* (hypernym/hyponym realtion)of the target wordnet were considered to select the equivalent mappings:

- *isEquivalent* (isCorrect): the correct equivalent mappings appear among the top k candidate synsets.
- *isHypernym*: the candidate synset is a hyponym of the correct mapping.
- *hasHypernym* (or isHyponym): the hypernym of the candidate synset is an equivalent mapping.
- *isSister* : the candidate synset is a sister node of an equivalent mapping (has the same hypernym synset).

Figure 7.6 (middle) plots the upper and lower bounds, the precision, recall, and f-measure of the obtained mappings using the Google translator for the ArWN synsets.

| No.	Experiment	Mappings Provided	Correct Mappings provided	Coverage	Upper Precision	Upper Recall	Upper F-measure
			Mapping ArWN-to-EnWN				
1	Google Translation	9425	7889	90.59	83.70	75.83	79.57
2	Sina&Google Trans.	9523	8340	91.53	87.58	80.16	83.71
3	Oracle Ttranaslation	10350	10335	99.43	99.90	99.34	99.62
4	Extended Oracle Trans.	10350	10339	99.48	99.87	99.38	99.62
5	All Dictionaries	10350	10344	99.52	99.90	99.42	99.66
	Lower Bound (First Sense Heuristic) ; range [k=1 -to- k=100]						
13	Google Dict.				25.8-55.4	23.4 - 42.0	24.6 - 47.8
14	Sina&Google Dict.				26.4 - 48.3	24.2 - 44.2	25.2 - 46.2
15	Oracletranaslation				47.8 -60.8	47.6-60.7	47.9-61.0
16	Extended Oracle				48.9 -59.6	48.6-59.3	48.7-59.4
17	All Dictionaries				43.0 -60.4	42.7-60.0	42.9 - 60.2
			Mapping ItWN-to-EnWN				
18	Google Translation	23,707	20,572	92.66	86.77	80.4	83.5
19	Extended Oracle Trans.	25,584	25,584	100	100	100	100
	Lower Bound (First Sense Heuristic) ; range [k=1 -to- k=100]						
20	Google Translation				52.0-66.0	48.2-61.5	50.1 - 63.8
21	Extended Oracle Trans.				72.0 - 73.0	72.0 - 73.0	72.0 - 73.0
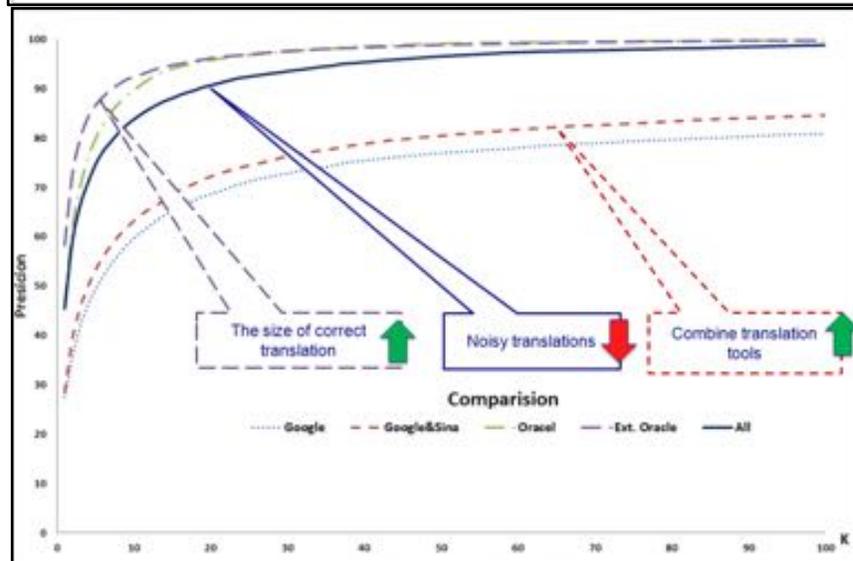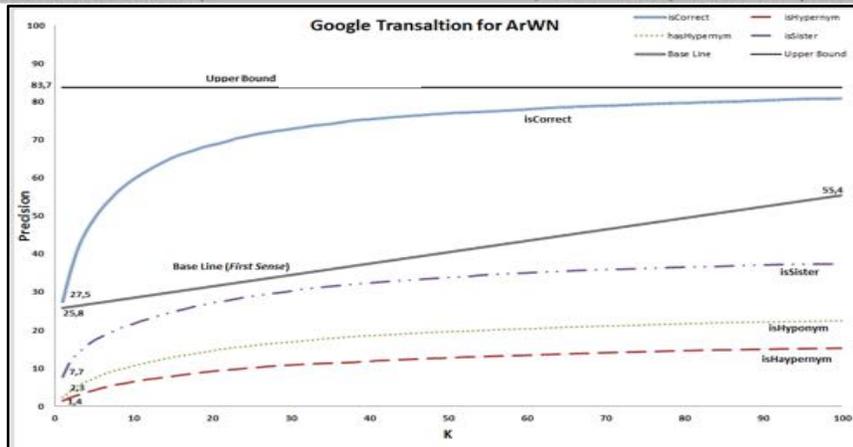
Figure 7.6 (Top) overall experiment results. (Middle) mappings using Google translation for ArWN. (Bottom) compression between the translation methods for ArWN

From the results in Figure 7 we can notice that, although the translation is important, providing an effective ranking of the correct senses is a crucial task; high coverage and precision with the translation is achieved only for high value of k. In fact, the goodness and the better performance of the mapping algorithm is to provide the correct mappings while minimizing the value of k, and this depend on the ranking function. If we consider the selection of top-1 concept, i.e., the best match is selected to be part of the alignment, the precision is still quite low.

Top-3 mappings found by the algorithm between the Arabic and English concepts can be found at https://www.dropbox.com/sh/1xxq5nr73rhhgmf/AACAWFWe6UW2OqcSvdjBvJEQa

The performed experiments have demonstrated several outcomes that can be summarized as follow:

- the approach was tested over two different pairs of languages, which demonstrates its adoptability across different languages.
- The proposed approach outperforms the baseline settings. The upper bound indicates that there is a space for more improvements in terms of obtaining the correct translations and to better rank the candidate senses.
- The upper and lower bound performance of the proposed approach suggest that a semi-automatic matching approach should be preferred in this context.
- Using structural information encoded in the target wordnet improves the sense selection task.
- NLP techniques (e.g., stemming, headword extraction, ..etc) are expected to improve the Machine Translation (MT) coverage, and obtain more candidate senses (instead of using pure translation-lookup and word-sense exact matching).
- Combining the translations of MT tools with a bilingual dictionary translation improves the results (see Figure 7, bottom).
- Features obtained from the MT tools (Google translation) such as the translation score and the synset translations need to be explored in order to filter the correct translations and to better rank the candidate senses.

7.5 Open Research Directions

7.5.1 Building reference alignments

We introduced a classification-based mapping for cross-language matching purposes. We plan to implement the experiment based on classification tasks assigned to bilingual speakers to map concepts in the Arabic Ontology to Core WordNet concepts. For this goal, we plan to adopt crowdsourcing methods [Sarasua et al. 2012] to collect users feedback so as to converge on a set of shared agreed mappings. Through this experiments we would also like to validate the language-dependence hypothesis of the salient (core) concepts.

7.5.2 Semi-automated creation of linguistic ontologies

We plan to investigate the extent to which cross-language matching methods can be exploited to support the (semi)-automated creation of large linguistic ontologies. For this, mappings between unstructured or weakly structured lexical resources (e.g., collections of synsets) and structured ontologies like WordNet can be found using a mapping algorithm as the one described in this deliverable. After the mappings are established, new relations between

unstructured lexical elements can be derived from the relations between the concepts in the structured ontology.

The experiment based on classification tasks assigned to bilingual speakers requires a large number of human inputs and can be practically difficult at large scale (e.g., for mapping several thousands of concepts). We therefore plan to use a different CL-WSD task: given a synset in one language the mapping algorithm is used to present to users a set of top-k WordNet concepts matching the synset; the users classify the synset as equivalent to one of the suggested WordNet concept.

With respect to the matching methods used in the algorithm we plan to investigate the use of Explicit Semantic Analysis (ESA) in our matching problem [Gabrilovich and Markovitch 2007, Sorg and Cimiano 2012] so as to enhance the word sense selection (conceptual translation) task. This idea is based on recent work at UNIMIB, where ESA has been applied to match short descriptions of domain entities. Moreover, we plan to explore different NLP techniques (e.g., stemming, headword extraction, ..etc) to improve the Machine Translation (MT) coverage, and to obtain more candidate senses (instead of using pure translation-lookup and word-sense exact matching).

With respect to inference of relations between concepts, we could to define a different mapping weight in the above mentioned *CL-WSD* task. Mainly, we need to represent an equivalent mappings stating from subsumption relation by considering the mappings weight (confidence level). As a subsequent step, we need to define and develop algorithms for semantic relations inference and to validate such methods using the cross-language mappings gold standard.

Finally, following the encouraging results obtained by the proposed mapping algorithm, we plan to investigate the construction of partial structural source synsets (instead of mapping unstructured synsets), and to investigate its impact on the mapping algorithm inspired by the work presented in [Pilehvar and R. Navigli 2014].

REFERENCES

1. Mustafa Jarrar: Building A Formal Arabic Ontology. In proceedings of the Experts Meeting On Arabic Ontologies And Semantic Networks. Alecso, Arab League. Tunis, July 26-28, 2011.
2. Mustafa Jarrar: Arabic Ontology, Lecture Notes. Sina Institute, Birzeit University, 2012.
3. Mustafa Jarrar, Rana Rishmawi, Hiba Olwan: Top Levels of the Arabic Ontology. Technical Report. Sina Institute, Birzeit University, 2013.
4. Mustafa Jarrar: Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In Proceedings of the 15th international conference on World Wide Web, 2006.
5. Mustafa Jarrar: Lexical Semantics and Multilingualism. Lecture Notes, Sina Institute, Birzeit University, 2014.
6. Manuel Atencia, Alexander Borgida, Jerome Euzenat, Chiara Ghidini, and Luciano Serafini. A formal semantics for weighted ontology mappings. In International Semantic Web Conference, pages 17-33, 2012.
7. Paolo Bouquet, Luciano Serafini, and Mario Zanobini. Semantic coordination in systems of autonomous agents: the approach and an implementation. In WOA, pages 179-186, 2003.
8. Gerard de Melo and Gerhard Weikum. Constructing and utilizing wordnets using statistical methods. Language Resources and Evaluation, 46(2):287-311, 2012.
9. Cassia Trojahn dos Santos, Paulo Quaresma, and Renata Vieira. An api for multi-lingual ontology matching.In LREC, 2010.
10. Jerome Euzenat and PavelShvaiko. Ontology matching. Springer, 2007.
11. Jerome Euzenat. Algebras of ontology alignment relations. In International Semantic Web Conference, pages 387-402, 2008.
12. Christiane Fellbaum. Wordnet: An electronic lexical database. Cambridge, MA. MIT Press,1998.
13. Miller and Fellbaum, WordNet. 1991

14. Bo Fu, Rob Brennan, and Declan O'Sullivan. A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. J. Web Sem., 15:15-36, 2012.

15. Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asuncion Gomez-Perez, Paul Buitelaar, and John McCrae. Challenges for the multilingual web of data. J. Web Sem., 11:63-71, 2012.

16. Ernesto Jimenez-Ruiz, Bernardo Cuenca Grau, and Ian Horrocks. Logmap results for oaei 2012, 2012.

17. Jason J. Jung. Ontological framework based on contextual mediation for collaborative information retrieval. Inf. Retr., 10(2), April 2007.

18. Jason J. Jung, Anne Hakansson, and Ronald L. Hartung. Indirect alignment between multilingual ontologies:A case study of korean and swedish ontologies. In KES-AMSTA, pages 233-241, 2009.

19. Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell., 193:217-250, 2012.

20. Pavel Shvaiko and Jerome Euzenat. Ontology matching: State of the art and future challenges. IEEE Trans. Knowl. Data Eng., 25(1):158-176, 2013.

21. Dennis Spohr, Laura Hollink, and Philipp Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In International Semantic Web Conference (1), pages 665-680, 2011.

22. S. M. Harabagiu, editor. Proc. Workshop Usage of WordNet in Natural Language Processing Systems. ACL, Université de Mont´real, Mont´real, QC, Canada, 1998.

23. Z. Gong, C. W. Cheang, and L. H. U. Web query expansion by WordNet. In Proc. DEXA 2005, Copenhagen, Denmark, volume 3588 of LNCS, pages 166–175. Springer, 2005.

24. Mamoun Abu Helou, Matteo Palmonari, Mustaf Jarrar, Christine Fellbaum. Towards Building Linguistic Ontology via Cross-Language Matching. The 7th Conference on Global WordNet, in Tartu (Estonia), January, 2014a.

25. Mamoun Abu Helou, Matteo Palmonari. Extreme Cross-Language Matching for the Creation of Large Linguistic Ontologies: Mapping framework and preliminary experimental analysis, (submitted) OM-ISWC 2014b.

26. S. Hertling and H. Paulheim. WikiMatch - Using Wikipedia for Ontology Matching. In Proceedings OM, 2012.

27. Liang, A., Sini, M.: Mapping AGROVOC \& the Chinese Agricultural Thesaurus: Definitions, Tools Procedures. New Review of Hypermedia \& Multimedia, 2006.\bibitem{hirst} Hirst G.. Ontology and the Lexicon, in Handbook on Ontologies and Information Systems. eds. S. Staab and R. Studer. Heidelberg: Springer, 2004.R.

28. Navigli. Word Sense Disambiguation: a Survey. ACM Computing Surveys, 2009.

29. Samer Hassan and Rada Mihalcea, Cross-lingual Relatedness using Encyclo pedic Knowledge, to appear in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), 2009.

30. Ngai, G., Carpuat, M., Fung, P.: Identifying Concepts Across Languages: A First Step towards A Corpus-based Approach to Automatic Ontology Alignment. In: Proceedings of the 19th International Conference on Computational Linguistics, vol. 1, pp. 17 (2002)

31. M. T. Pilehvar and R. Navigli. A Robust Approach to Aligning Heteroge neous Lexical Resources. Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June 22-27, 2014

32. Venetis, Petros and Garcia-Molina, Hector and Huang, Kerui and Polyzotis, Neoklis. Max Algorithms in Crowdsourcing Environments. Proc. WWW 2012.

33. Isabel F. Cruz, Alessio Fabiani, Federico Caimi, Cosmin Stroe, and Matteo Palmonari. Automatic configuration selection using ontology matching task profiling. In ESWC, pages 179-194, 2012.

34. Narducci F., Palmonari M, Semeraro G.. 2013. Cross-language Semantic Retrieval and Linking of E-gov Services. he 12th ISWC and the 1st Australasian Semantic Web Conference , October, Australia

35. Gabrilovich, E. and Markovitch, S. 2007. Computing semantic relatedness using wikipediabased explicit semantic analysis. In Proceedings of the 20th IJCAI'07, pp1606–1611, San Francisco, CA, USA.

36. Sarasua C., Simperl E., Noy N.F.2012. CROWDMAP :Crowdsourcing Ontology Alignment with Microtasks. In ISWC-2012,. Springer.

8 Quality Control and Self Evaluation

The writing of this deliverable went through several iterations. To evaluate the quality of the pre-final version, according to the quality control procedure defined in WP5 (see task 5.3), some partners were assigned by the coordinator to review the deliverable and provide their feedback. Since all partners have contributed to this deliverable, reviewers were assigned sections that didn't author: MICHAEL (section 3) was reviewed by Amanda Hicks-BBAW; KYOTO (section 4) was reviewed by Rute Costa -UNL; OKKAM (section 5) was reviewed by Matteo Palmonari-BICCOCA; Organic.EduNet (Section 6) was reviewed by Christophe Roche-UNL; and the Mapping Framework (section 7) was reviewed by Stefano Bortoli-UNITN. The coordinator, Mustafa Jarrar-BZU, reviewed the overall quality and ensured that all feedback received from the reviewers was reflected in the final version of the deliverable.