SIERA

Integrating Sina Institute into the European Research Area

FP7- 295006

D2.1

Intermediate Report on Research Setup

Document Identifier	SIERA FP7-INCO-2011-6 (295006) /2013/D2.1
Version	Version 1.0
Date	29 May 2013
State	Final
Distribution	**PP =** Restricted to other programme participants (including the Commission Services)

Executive Summary

The core objective of the research setup activities defined in WP2 is to set up close research cooperation between Sina Institute and EU partners, which will facilitate and enable the exchange of knowledge and enhancement of the cooperation capacity of Sina BZU . The expected result is the establishment of a solid research setup that can enable the production of future research work.

Four different activities took place to achieve the above-mentioned objective, which will be discussed further in section two. In order to ensure the effective performance of those activities, they were based on the use of two multilingual knowledge sharing portals (MICHAEL and KYOTO). These portals are necessary to investigate the enabling and integration of Arabic language and content into the portals, as well as to establish a mapping framework that maps the Arabic Ontology to a widely adopted lexical ontology, such as WordNet.

The activities and their main achievements are as follows :

- **Investigate the support of BZU Sina Arabic processing APIS:** The missing Arabic tools were specified and were designed and finalized to work with minimal installation on the MICHAEL search system, that is they are ready for demo testing. The activity results which currently depends on the results of integration process, will be provided by MICHAEL developers after testing the tools.
- **Extending two EU multilingual portals**: MICHAEL's thesaurus was extended into Arabic.
- **Establishing a framework for mapping between WordNet and Arabic Ontology**: SIERA partners will defined the mapping framework and extend KYOTO ontology to Arabic by mapping the Arabic Ontology Top Levels.
- **Resolving and linking Arabic entities with existing entities in MICHAEL and KYOTO :**The Arabic cultural objects was defined and yet to be connected to MICHAEL; the Bethlehem Thesaurus SKOS format was produced, using the TMP environment, after defining a domain ontology for Bethlehem data. Furthermore, Bethlehem's historical buildings and Arabic named entities extracted from Wikipedia were OKKAM-ized.

Document Information

Project Number	FP7–295006	**Acronym**	SIERA
Full Title	Integrating Sina Institute into the European Research Area		
Project URL	Sina.birzeit.edu/SIERA/		
Document URL	-		
EU Project Officer	Tanya Dimitrova		

Deliverable Number	2.1	**Title**	Intermediate Report on research setup
Work Package Number	2	**Title**	Networking and Research Setup

Date of Delivery	**Contractual**	30 Mar, 2012	**Actual**	28 May,2013
Status	version 1.0		final ×	
Nature	Prototype	Report ×	Dissemination	
Dissemination Level	Public	Consortium ×		

Authors (Partner)	Mamoun Abu Helou, Mustafa Jarrar , Matteo Palmonari, Ali Salhi, Amanda Hicks, Stefano Bortoli, Christophe Roche, Christiane Fellbaum, Paolo Bouquet, Adnan Yahya.			
Resp. Author	Mamoun Abu Helou (BZU)		**E-mail**	mabuhelou@Birzeit.edu
	Partner	BZU	**Phone**	+972(2)2982917

Document Abstract	The purpose of this deliverable is to report on the consortium's activities to set up research and cooperation between Sina institute and EU partners, in which exchanging knowledge and enhancing the cooperation capacity of BZU Sina will be facilitated. The report will cover all activities and outcomes of research and cooperation setup (Task 2.2) up to the month of delivery [month 18].
Keywords	Research Setup, Multilingual Portals, Mapping Framework

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
2013-02-3	v.1	Mamoun Abu Helou (BZU), Mustafa Jarrar (BZU)	Create and drafting
2013-03-23	v.2	Ali Salhi, Adnan Yahya (BZU)	Editing sections B.1-3
2013-04-11	v.3	Matteo Palmonari (UniMiB) Christiane Fellbaum (BBAW) Mustafa Jarrar (BZU)	Feedback on the Mapping Framework
2013-04-12	v.4	Amanda Hicks (BBAW)	Description of the Mapping Between the Arabic Ontology Top Levels and KOYOT
2013-04-14	v.5	Mamoun Abu Helou (BZU)	Edit and integrate the mapping framework
2013-04-16	v.6	Ali Salhi (BZU)	Edit sections B.1-3, introduction and conclusion.
2013-04-25	v.7	Mamoun Abu Helou (BZU)	Editing the mapping framework
2013-05-26	v.8	Christophe Roche (FCSH-UNL)	Feedback and quality control
2013-05-27	v.9	Stefano Bortoli , Paolo Bouquet (UniTN)	Description of the OKKAM-ization of the Arabic entities
2013-05-29	V1.0	Mamoun Abu Helou (BZU)	Integration, implementation of quality control comments, and final version

Project Consortium Information

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

Associate Partners

- Vrije Universiteit Amsterdam (Contact Person: Prof. Piek Vossen)
- Michael Culture Association (Contact Person: Ms. Marie-Véronique Leroi)
- Center of Cultural Heritage Preservation – Bethlehem (Contact Person: Ms. Nada Atrash)
- Ministry of Telecom and Information Technology
- Engineering Company for the Development of Digital Systems (RDI) (Contact Person: Dr. Mohsen Rashwan).

Work Package Structure

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	BZU	37.00
2	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	

A. Introduction

The core objective of the research setup activities demonstrated inWP2 is to set up close research cooperation between Sina institute and EU partners, in which exchanging knowledge and enhancing the cooperation capacity of BZU Sina will be facilitated. To achieve this goal, different activities took place with the support of two multilingual knowledge sharing portals (MICHAEL and KYOTO) to investigate the enabling and integration of Arabic language and content into the portals. The aim of these activities is to enable close and sustainable scientific cooperation between EU scientists and BZU Sina Institute while integrating their portals and tools, and/or to identify the missing components that need further research and development.

Four activities took place in this regarding , which will be discussed in section two , where we (in section B.1) will investigate the support of Sina-BZU Arabic processing APIS in multilingual knowledge sharing portals and highlight BZU tools that can be used. We will also report on what has been achieved so far (Activity One), also we will discuss the progress we have done so far in extending two EU multilingual knowledge sharing portals (MICHAEL and KYOTO) in section B.2 (Activity Two), we will discuss our work on resolving and linking Arabic entities with entities in MICHAEL and KYOTO in section B.3 (Activity Three), and finally we will discuss about the framework used for mapping between WordNet and Arabic Ontology in section B.4 (Activity Four).

In order to move forward in the planned activities we need first to introduce MICHAEL and KYOTO portals.

a) MICHAEL Portal:

MICHAEL - 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 European Commission'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 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 labelling (see Figure 1).

b) KYOTO portal

KYOTO[2] project 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, similar to the KYOTO ontology that was constructed by Amanda Hicks(BBAW), as an extension of DOLCE-DNS Ontology, the KYOTO project also includes ontology-lexicon mapping tools to WordNet, which is free English Lexical Database.

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

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

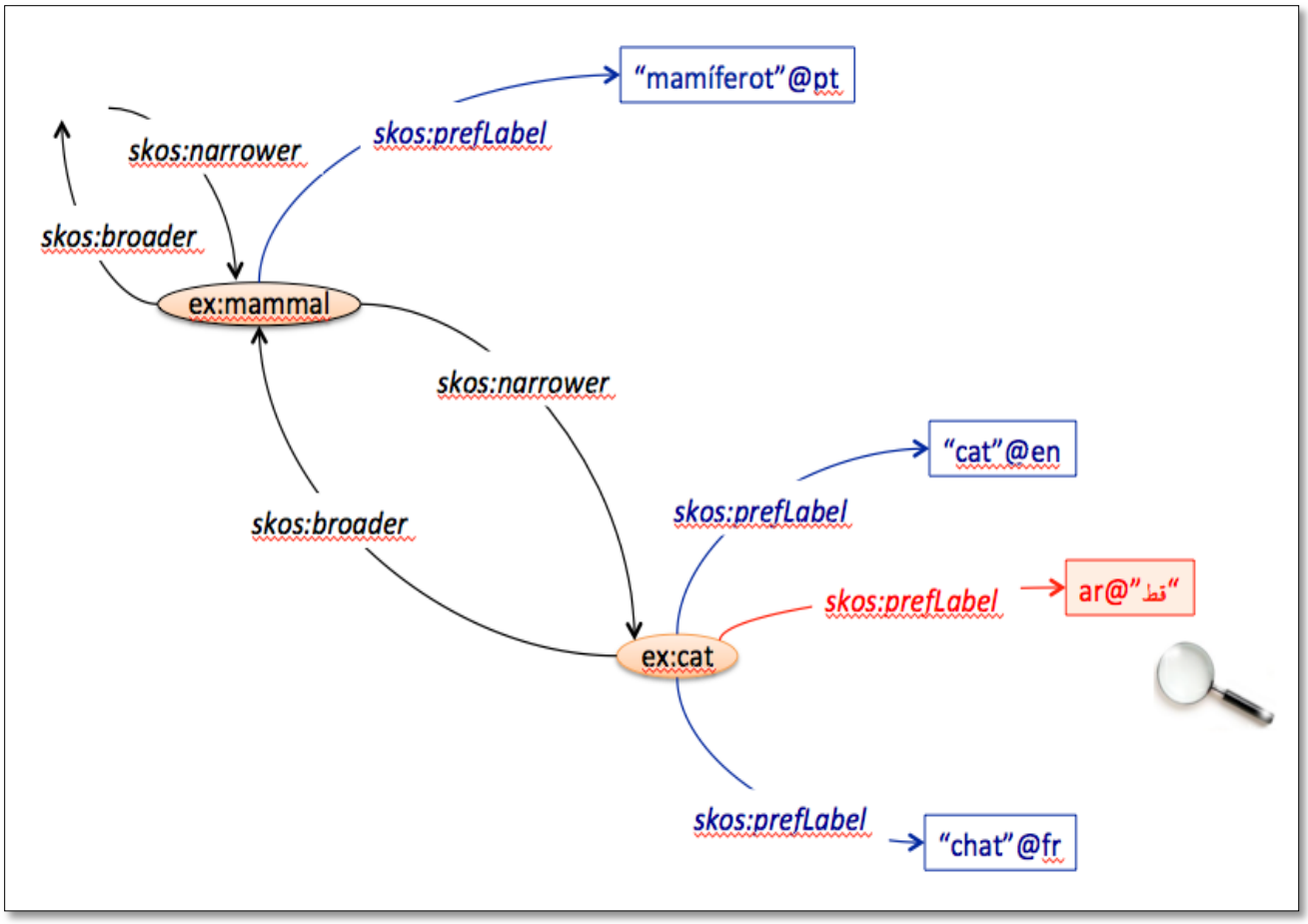

Figure 1: An excerpt of Michael thesaurus SKOS representation. the Arabic label "ar@قط" for ex:*cat* shown in red

B. Planned Activities and Research Results

1. Investigating Sina-BZU Arabic processing APIs

1.1. Participants: BZU, FCSH-UNL, BBAW, MICHAEL & KYOTO

1.2. Description :

Investigating Arabic support in multilingual knowledge sharing portals:

BZU Sina has developed several Arabic processing APIs (e.g., stemmer, root extractor, language detector, query expander, proper names checkers/translators, and others), which are basic building blocks for Arabic search engines. In this activity we will investigate the support of Sina-BZU Arabic processing APIS in multilingual knowledge sharing portals (MICHAEL and KYOTO) and highlight BZU tools that will be used and investigated , as a result of this the know-how on building multilingual knowledge sharing portals should be transferred to BZU Sina, and the missing APIs or those that need further research should be identified. To achieve this: (i) a workshop will be organized for the partners to demonstrate their in-house tools in details; (ii) EU partners will give direct access, or a snapshot of the source code, to BZU Sina for experimental purposes; (iii) BZU Sina will collect an experimental sample of 1000 objects related to Arabic culture and ecology and upload it to MICHAEL and KYOTO; (iv) BZU Sina, with active support and involvement of EU partners, will tune its tools and APIs with MICHAEL and KYOTO. The output of this activity will be a report specifying the missing APIs or those that need further research and development, which can be jointly tackled by the project partners in the future. FCSH-UNL, BBAW and BZU Sina, and possibly with some support from the project associates, will visit and cooperate with each other in order to carry out this activity.

1.3. Results:

The partners demonstrated their in-house tools in details during SIERA Kick-off Conference (24/11/2011) ,where different discussions and presentations took place including the presentation of MICHAEL and KYOTO [3]. SIERA partners had also attended a workshop a day earlier (20/9/2012) in Paris , the workshop was dedicated to the terminology management platform (TMP) – a part of FP7 Linked Heritage project. Moreover, the SIERA partners have conducted a technical meeting in Paris (21/9/2012), and the meeting was also dedicated to "TPM (Terminology Management Platform) – ;however, focusing on Extending MICHAEL thesaurus with Arabic & to investigate the needed and missing Arabic APIs and tools".

In this meeting, Adnan Yahya & Ali Salhi (BZU) presented the available tools and how they can be used and integrated into working systems and search engines, followed by a technical discussion with the participant of Christophe Roche (CLUNL), and Marie-Véronique & Florent André (TMP & MICHAEL) to discuss the integration of the tools in MICHAEL. Later, several online discussions & emails carried out between involved partners in this task to investigate the needed and missing Arabic APIs and tools.

It was agreed that the missing tools that need to be included and tested on MICHAEL search engine are the following:

- **Arabic Language Detector:** A tool that detects the Arabic documents. The tool returns true if the input document is written in Arabic and false otherwise.
- **Arabic Spell checking 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 provided is 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.

[3] For more details about the Kick-Off meeting and presentations (available online), please check: http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-kick-off-conference/

Adnan Yahya & Ali Salhi (BZU) designed and finalized the above tools to work with minimal installation and provided them as standalone services that can be easily integrated with any search engine built with Java programming language; including the case of MICHAEL Search engine which is based on Apache Lucene/Solr an open source enterprise search platform written in Java. The tools are ready for demo testing and was tuned for the experiment of integrating them with MICHAEL search engine. The tools were sent to MICHAEL developers with instructions of how to be integrated in the query processing system of MICHAEL search engine.

The missing Arabic tools were specified and were designed and finalized to work with minimal installation on MICHAEL search system and is ready for demo testing. The activity results currently depends on the results of integration which will be provided by MICHAEL developers after testing the tools. Please note that the tools that need further research and development will be highlighted after the conduction of the integration test, those tools can be jointly tackled by the project partners in the future.

2. Extending two EU multilingual portals

2.1. Participants: BZU, FCSH-UNL, BBAW, BICOCCA.

2.2. Description

This activity complements the previous activity, investigating Sina-BZU Arabic APIs; that is, the multilingual support in knowledge sharing portals, such as MICHAEL and KYOTO, is typically based on multilingual thesauri/ontologies that represent domain concepts. Thus, to enable Arabic in such portals, the MICHAEL's multilingual thesaurus and KYOTO's multilingual ontology need to be extended to support Arabic concepts. This process requires know-how in formal knowledge representation as well as multilingual lexical semantics. Partner's methodologies, tools, and know-how will be utilized to help BZU Sina carry out this activity.
In particular, (i) a workshop will be organized for partners to demonstrate their in-house methodologies, tools, and resources, which include the FCSH-UNL's Onto Terminology methodology and tools to formally map between multilingual terms (languages level) and domain concepts (conceptual level). This methodology has been applied in several FP7 projects including Linked Heritage, Europeana, ASTECH, Athena, and others. For quality and evaluation, BICOCCA's methodology and tools will be used, as it has been applied in developing and evaluating the ONTO_PA Italian multilingual public administration ontology, which contains cultural and environment domain concepts; (ii) BZU Sina (with active guidance from FCSH-UNL and BBAW) will use the OntoTerminology to extend MICHAEL's multilingual thesaurus and KYOTO's multilingual ontology with Arabic concepts; (iii) The new concepts that may emerge from Arabic culture will be mapped into their equivalent Italian concepts by the BICOCCA, and into French and Portuguese by FCSH-UNL, with the involvement of BZU Sina. (iv) BICOCCA and BZU Sina will cooperate to evaluate the quality of the resultant Arabic extension and their mappings into other languages. The result of this activity will be an Arabic extension to MICHAEL's multilingual thesaurus and KYOTO's multilingual ontology. This extension is necessary for the previous activity, and helps BZU Sina enhance its skills in integrating multilingual databases and evaluate their quality, and most importantly, in preparing for future joint research and cooperation on multilingual lexical semantics. FCSH-UNL, BBAW, BICOCCA and BZU Sina researchers will visit and cooperate with each other to carry out this activity.

2.3. **Results:**

The results related to extending two EU multilingual portals were achieved throughout the participation of conferences, meetings and discussions**.** In the kickoff meeting (24/11/2011), Prof. Christophe Roche (CLUNL) presented and discussed UNL's Onto Terminology methodology and tools. Later**,** SIERA partners conducted a technical meeting and attended a workshop in Paris (20-21/9/2012), which were dedicated to "TMP (Terminology Management Platform) – Extending MICHAEL thesaurus with Arabic & to investigate the needed and missing Arabic APIs and tools". Many technical and scientific issues related to extending MICHAEL with Arabic were discussed, such as how to extend MICHAEL's functionality to enable search in Arabic (which we covered in section II.1; Investigating Sina-BZU Arabic processing

APIs), how to extend MICHAEL with Arabic entities (which will be discussed in section II.4; resolving and linking Arabic entities with entities in MICHAEL and KYOTO and how to extend MICHAEL's Thesaurus to include Arabic. Regarding extending MICHAEL's Thesaurus into Arabic, Christophe Roche (CLUNL), and Marie-Véronique & Florent André (TMP & MICHAEL) introduced & presented the technical details of MICHAEL . Technically speaking MICHAEL can be extended using TMP (Terminology Management Platform), MICHAEL depends on SKOS interchange format. Such a format clearly separates the conceptual dimension of a thesaurus from its linguistic dimension, following in this way the OntoTerminology methodology. Therefore, the Arabic localization of the MICHAEL thesaurus relies on the preferred and alternative lexical labels with the Arabic language tag, and extending the MICHAEL thesaurus with Arabic content relies on the SKOS semantic relationships. Such an approach requires a skosification of thesauri. To this end, a strong cooperation with the Terminology of the FP7 Linked Heritage has been set up.

In order to progress further in extending MICHAEL, a practical training session was carried out by the TMP & MICHAEL team on how to use TMP to extend MICHAEL with Arabic took place in the technical meeting in Paris. Such practical training will provide the partners with the knowledge and skills needed to achieve and successfully implement this activity. To achieve this activity goal (extending MICHAL with Arabic) , MICHAEL's thesaurus in SKOS format was extended with Arabic by Mustafa Jarrar & Rana Rishmawi (BZU) and sent back to Christophe Roche (CLUNL) for quality control and feedback. Figure 2, depicts the "Educational sciences and environment" concept with its different labels in different languages including the Arabic in red.

The process and activities implemented in extending the multilingual ontology KYOTO to Arabic will be addressed in the next section .

```
<rdf:Description
rdf:about="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Educational
_sciences_and_environment">
   <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
   <skos:inScheme
rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects"/>

   <skos:prefLabel xml:lang="cs">Vědy o vzdělávání a vzdělávací prostředí</skos:prefLabel>
   <!-- <skos:prefLabel xml:lang="ee">Haridusteadused ja keskkond</skos:prefLabel> -->
   <skos:prefLabel xml:lang="en">Educational sciences and environment</skos:prefLabel>
   <skos:prefLabel xml:lang="it">Scienze dell'educazione</skos:prefLabel>
   <skos:prefLabel xml:lang="fr">Sciences de l'éducation et milieu éducatif</skos:prefLabel>
   <skos:prefLabel xml:lang="fi">Kasvatustieteet ja -ympäristö</skos:prefLabel>
   <skos:prefLabel xml:lang="sv">Pedagogik</skos:prefLabel>
   <skos:prefLabel xml:lang="el">Παιδαγωγική και εκπαιδευτικό περιβάλλον</skos:prefLabel>
   <skos:prefLabel xml:lang="nl">Onderwijswetenschappen</skos:prefLabel>
   <skos:prefLabel xml:lang="lv">Izglītības zinātne</skos:prefLabel>
   <skos:prefLabel xml:lang="hu">Oktatástudomány és környezet</skos:prefLabel>
   <skos:prefLabel xml:lang="bg">Оразование и среда</skos:prefLabel>
   <skos:prefLabel xml:lang="pl">Dydaktyka i środowisko nauczania</skos:prefLabel>
   <skos:prefLabel xml:lang="sk">Vedy o vzdelávaní a vzdelávacie prostredie</skos:prefLabel>
   <skos:prefLabel xml:lang="es">Ciencias de la educación y ambiente educacional</skos:prefLabel>
   <skos:prefLabel xml:lang="ar">تربوية وبيئة علوم</skos:prefLabel>

   <skos:narrower
rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Educatio
n"/>
   <skos:narrower
rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Learning"
/>
   <skos:narrower
rdf:resource="http://www.athenaeurope.org/athenawiki/AthenaThesaurus/Michael_Subjects#Educatio
nal_history"/>

 </rdf:Description>
```

Figure 2: an excerpt of the extended SKOS file.

3. Resolving and linking Arabic entities with entities in MICHAEL and KYOTO

3.1. Participants: BZU, UNTIN, FCSH-UNL, BICOCCA.

3.2. Description :

When resolving and linking entities and identities, one must keep in mind that most entities (e.g., people, organizations, products, places, events) have different names in different languages, which is a major challenge faced when integrating knowledge from different sources, cultures, and languages. This preparatory activity aims to investigate disambiguating Arabic entities and linking them with entities in MICHAEL and KYOTO. The OKKAM entity management system (developed by UNITN) will be used to facilitate the disambiguation and entity linkage. The methodology to assess the quality of entities and concepts (developed by BICOCCA) will be also used to control the evolution of the linked data and its ontology.

In this activity BZU Sina, UNITN, and FCSH-UNL will cooperate, and use the OKKAM functionalities, to map between the Arabic cultural heritage and ecology entities and the related entities in MICHAEL and KYOTO portals. The name correction/translation tools developed at BZU Sina will be used to facilitate this investigation process. BZU Sina and BICOCCA will cooperate to validate the quality of these linked entities, using BICOCCA's methodology and tools. This methodology will be also used to manage the evolution of linked data, and other sources, that are characterized by different time stamps and histories. The strength of BICOCCA's methodology lies in the innovative algorithms it uses for temporal and spatial record linkage and the co-referencing algorithms for matching of concepts hidden in linked data with the concepts of the ontology. The main outcome of this preparatory activity will not only be resolving and linking Arabic entities and names with MICHAEL and KYOTO, but also exchanging know-how and skills on entity disambiguation, quality, and evolution of linked data, as well. This activity will be carried out by UNITN, FCSH-UNL, BICOCCA, BZU Sina, and maybe with the support of the project associates. It might be worth noting that the cooperation in this activity will be exposed to further research and cooperation in an ongoing initiative that uses OKKAM for integrating huge resources of news articles, where Arabic content will be also used.

3.3. Results:

As part of this activity SIERA partners conducted a technical meeting and attended a workshop in Paris (20-21/9/2012), the meeting was dedicated to "TMP – Extending MICHAEL thesaurus with Arabic & to investigate the needed and missing Arabic APIs and tools".

In the meeting several issues related to extending MICHAEL with Arabic were discussed, such as how to extend MICHAEL's functionality to allow searching in Arabic (which we covered in section B.1; Investigating Sina-BZU Arabic processing APIs), how to extend MICHAEL's Thesaurus with Arabic (which we covered in section B.2) and how to extend MICHAEL with Arabic content.

Regarding the Arabic content, it was agreed that MICHAEL will be enriched with an experimental sample of 1000 objects from Bethlehem City related to Arabic culture through CCHP (Centre for Cultural Heritage Preservation) in Bethlehem. Also MICHEAL will be extended with another experimental sample of about 3000 named entities (about famous/old people) which was processed and extracted from Arabic Wikipedia by BZU. Next, we discuss the result of the SKOSfication of the Bethlehem Thesaurus and the OKKAMization of the Arabic entities.

SKOSfication of the Bethlehem Thesaurus

As we mentioned before, the MICHAEL Thesaurus is written in SKOS. Therefore, the Bethlehem Thesaurus must be SKOSified. The Bethlehem Thesaurus is about the historic town of Bethlehem. Buildings are described as a set of attributes with values (around 1000 building, each of them described by 36 attributes), a more artificial intelligence-oriented description than thesaurus-oriented. For example, type of construction:

Type of Construction
☐ 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)

Therefore, the SKOSification of the Bethlehem Thesaurus consists of defining a controlled and structured vocabulary from the values of the attributes, to do so the task was divided into twofold.
The first was to define the domain ontology using an ontology editor (SKOS is not a modelling language). Figure 3 represents one of the views of this ontology. 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.
Then, the SKOSification of the thesaurus was done using the TMP environment. TMP, for Terminology Management Platform, is an outcome of the FP7 Linked Heritage Project. It allows to define terminology in SKOS format.

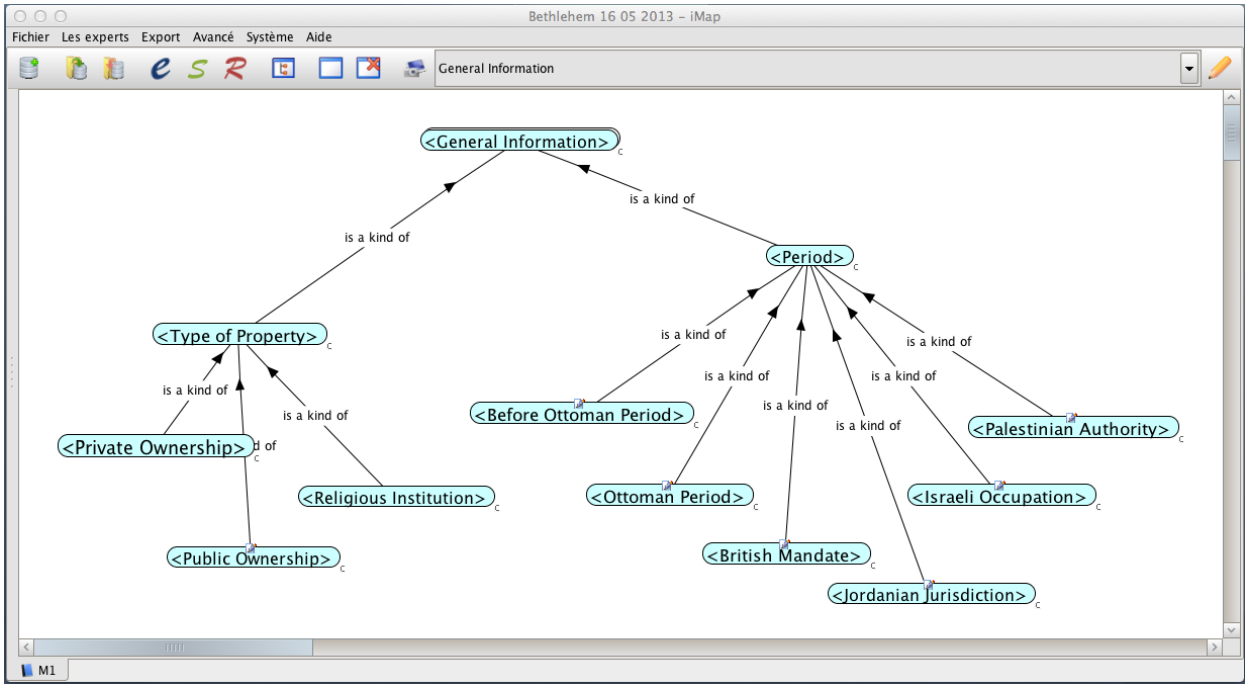

Figure 3 : An excerpt of Bethlehem SKOS ontology

Resolving and linking Arabic entities

The OKKAM entity management system (developed by UNITN) will be used to facilitate the disambiguation and entity linkage between experimental samples and MICHEAL and to do so , several meetings took place between Ali Salhi, Rana Rishmawi and Samer Zain (BZU), and CCHP (Centre for Cultural \Heritage Preservation) in Bethlehem, some with the attendance of Paolo Bouquet, and Stefano Bortoli (UNITN - OKKAM). The output of those meetings was the providing and preparing of 1000 GIS-enabled cultural objects to be included in OKKAM which will be linked to the MICHAEL multilingual portal. The RDF version of the objects was produced by Mamoun Abu Helou (BZU) . Also regarding the activity goal , Adnan Yahya & Ali Salhi (BZU) provided an experimental sample of 3000 named entities

(about famous/old people) extracted from Arabic Wikipedia to be included in OKKAM and then linked with MICHAEL multilingual portal.

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. The okkamization process consists in minting and maintaining a globally unique and persistent identifier through the Okkam Entity Name System APIs[4]. Each identifier is tied with an entity profile, and a set of alternative identifiers. In particular, the entity profile[5] is used to support the execution of sophisticated entity matching algorithms, therefore enabling reuse of the reuse of the minted identifiers. In fact, once an identifier is minted in the Entity Name System, third parties can attempt retrieving 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[6]. Using the latest, users can submit identification request using the Entity Identifier Request Language[7] , and lookup the identifier of the entity of interest .

So far, two datasets of Arabic entities were okkamized:

- the historical buildings of Bethlehem
- named entities extracted from Arabic Wikipedia[8]

The first dataset was built processing data collected by the Centre for Cultural Heritage Preservation (CCHP) of Bethlehem, that produced a census of the historical building to be preserved. In all, 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 the their identifiers particularly complicated. Therefore, we decided to postpone the okkamization of the remainder of the historical buildings. The descriptions of the historical buildings in Bethlehem included geospatial information that allowed us to place them on the map of entities Okkam is maintaining. A view of such map with the detail of one of the buildings is presented in Figure 4.

The second dataset was built crossing the descriptions produced by the Birzeit University team processing the Arabic Wikipedia, with the English DBPedia. This operation was done relying on the Open Refine tool. The Dbpedia entities were integrated with the Arabic names of entities obtained processing the Arabic Wikipedia, to produce descriptions that could be okkamized. This processed allowed us to create 1107 new entity profiles integrating Arabic names for entities. This is just a fraction of the 3000 extracted from Arabic Wikipedia. An example of the okkamization process executed is presented in Figure 5. The integration of Arabic names into the person profile allowed us to retrieve identifiers for such entities also using Arabic names, as showed in Figur 6. In fact, thanks to the extension of the descriptions with Arabic names, now entity identifies can be retrieved also using Arabic words.

[4] http://api.okkam.org
[5] a set of attributes in the form of (key, value) pairs
[6] http://api.okkam.org/search/
[7] http://project.okkam.org/intranet/entity-id-request-language
[8] http://ar.wikipedia.org/

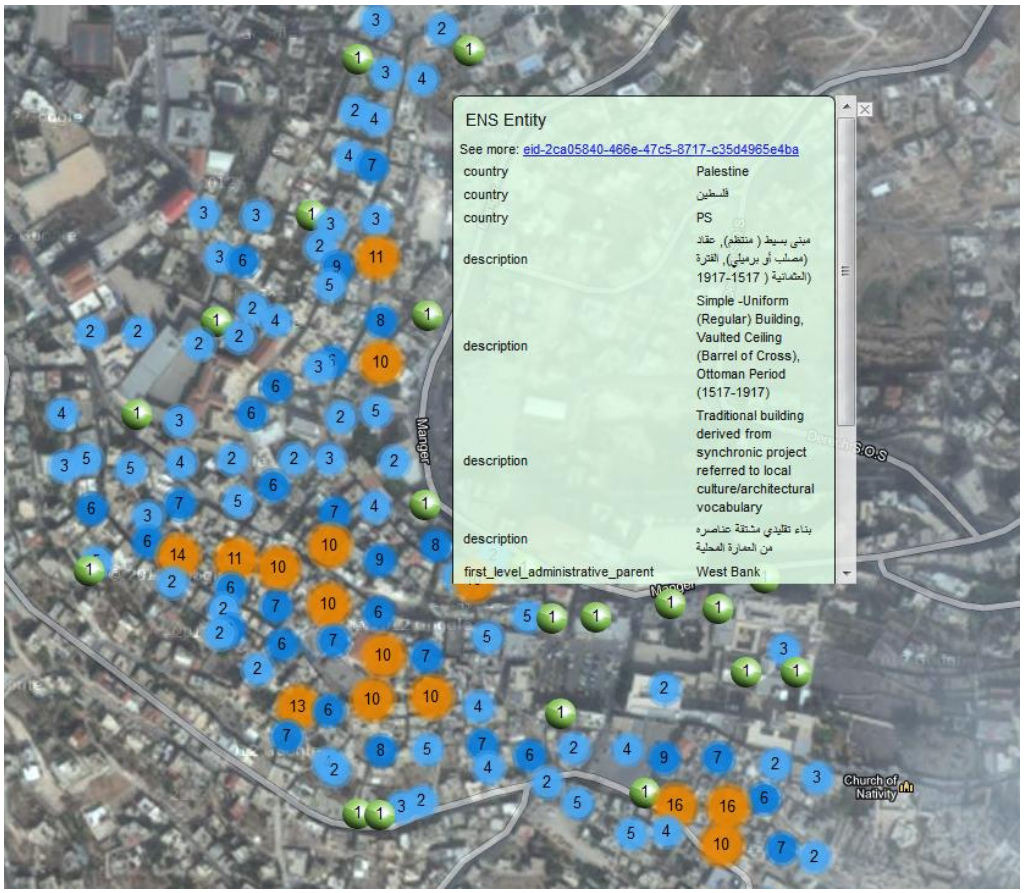

Figure 4: A view of the map of Bethlehem with the details of one of the building okkamized.

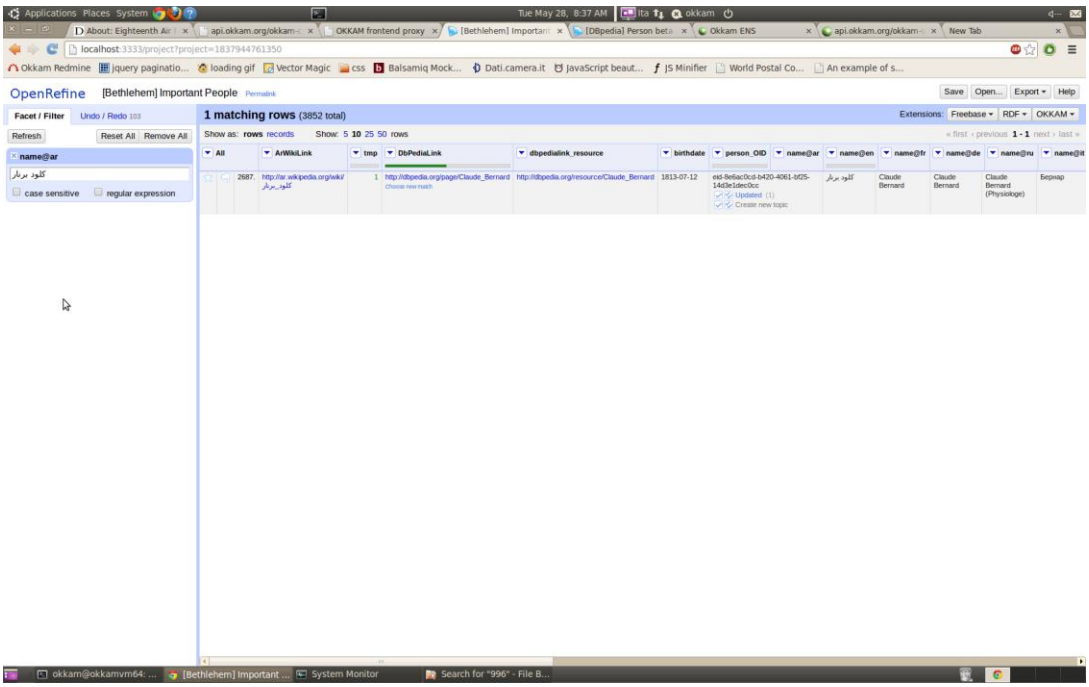

Figure 5: A screenshot of Open Refine processing a DBpedia record including the Arabic Names

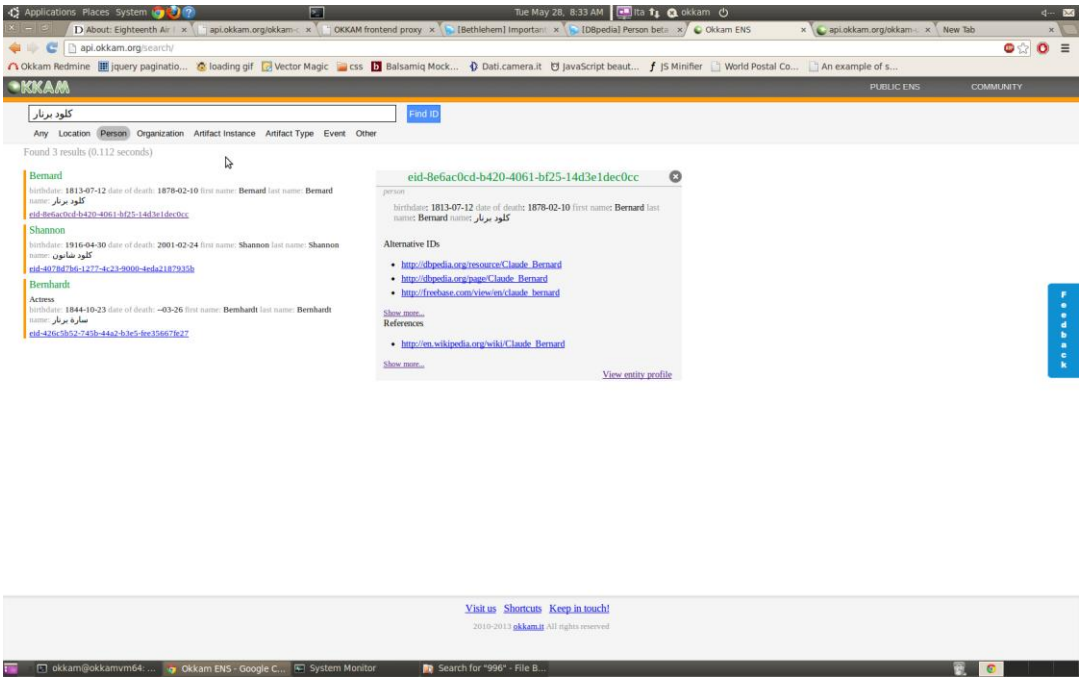

Figure 6: A screenshot of Entity Name System Lookup Interface searching using Arabic Names.

4. Establishing a framework for mapping between WordNet and Arabic Ontology

4.1. **Participants:** BZU, BICOCCA, BBAW, UNTIN,FCSH-UNL.

4.2. **Description:**

This activity focuses on general multilingual terminologies (i.e., language-level), unlike the previous activity which focuses on domain specific (cultural heritage) concepts. WordNet, a well-known lexical database for English, will be mapped into the Arabic Ontology, which is a lexical database for Arabic currently being developed by BZU Sina. The full mapping between WordNet and Arabic Ontology is beyond the scope of this project and will be carried in future cooperation. However, a mapping framework will be established in this project, as a foundation step.

In this activity number of workshops will be organized for the partners to demonstrate their in-house lexical databases, know-how, and mapping tools; SIERA partners will cooperatively establish the mapping framework, which will formally describe how the mapping should be done, based on the OntoTerminology formal principles and tools. Also SIERA partners will extend the multilingual ontology KYOTO to Arabic.

4.3. **Results**

To achieve the planned objectives several meetings were carried out by SIERA partners. As mentioned before, SIERA partners demonstrated their in-house tools in details during SIERA Kick-off Conference. Several discussions and presentations took place, in particular; Prof. Christophe Roche (CLUNL) presented and discussed UNL's OntoTerminology methodology and tools. Prof. Christiane Fellbaum, (BBAW) discussed the knowledge representation for Concepts in lexical resources and abstract ontologies and Prof. Piek Vossen (Vrije Universiteit Amsterdam, Netherlands) presented KYOTO platform for sharing knowledge across cultures and languages. Prof. Carlo Batini, (UNIMIB) demonstrated the temporal and spatial co-referencing of linked entities and the quality assessment of the produced ontologies. More details can be found at http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-kick-off-conference/

In the Lexical Semantic and Cultural Heritage meeting in Trento (http://sina.birzeit.edu/SIERA/featured-news/news-events/siera-wp2-meeting-in-trento-italy/. July, 2012), SEIRA partners focused on launching a

framework for mapping between WordNet and Arabic Ontology, and how to extend KYOTO to Arabic. Along with several online discussions and emails the partners succeeded to establish a mapping between the Arabic ontology Top Levels and KYOTO, and to define a mapping framework to align the Arabic Ontology to WordNet. Next we provide more details concerning each activity.
Moreover, a team of junior researchers working on the mapping framework were sent to the University of Milano Biccoca (UNIMIB) in Milano, Italy (http://sina.birzeit.edu/news-and-events/a-team-from-sina-institute-visits-the-university-of-milano-bicocca-italy/ , 11-20/3/2012) to widen the junior researchers knowledge and improve their abilities in various topics. Professors and PhD researchers at Milano Bicocca University presented their knowledge and work to the visiting members, in various fields, especially in Ontology Matching.
In addition, Dr. Jarrar (BZU) had the chance to meet and network with EU experts in the ontology engineering field while attending International conference on Ontologies (Graz, July 2012). Fruitful discussions carried out about the ongoing activities of building ontologies through exploiting the knowledge stored in other linguistic ontolgies (e.g Wordnet). This meeting facilitated the establishment of new cooperation with other potential EU partners and the establishment of future research work.

Mapping Between the Arabic Ontology Top Levels and KYOTO

"*Top Levels of the Arabic Ontology*" provides a comprehensive overview of the meanings and formal and ontological properties of 63 concepts in the top levels of the Arabic Ontology (AO). The development of the top levels of AO aims to capture the most abstract concepts lexicalized in the Arabic language and model them according to the kind of ontologically rigorous criteria that are used to develop top level ontologies such as BFO, DOLCE, and SUMO. This is a novel approach to developing an electronic lexicon and ensures of combining the best aspects of an electronic lexicon and a more rigorous ontology.
This review focuses on the comparisons and links of AO with the KYOTO ontology with a particular emphasis on the definitions and descriptions of the terms. In short, of the 63 classes 45 of the comparisons between AO and KYOTO are extremely clear and accurate. Only 18 of the comparison stand in need of clarification or revision. This means 72% of the comparisons are accurate.
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. Every ontology is improved by successive revisions, and for a first release, Top Level Ontology AO represents an impressive effort. The majority of current comparisons of AO concepts with KYOTO deftly deal with abstract and technical issues accurately, and this is a significant achievement. What follows is simply a list of recommendations for rendering these comparisons even more clearly or precisely, but these recommendations do not diminish the significance of the AO.
What follows is a list of concerns regarding specific comparisons and links of AO with the KYOTO ontology.

- *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 insofar as both are root nodes, they are obviously comparable.
- *Abstract* – *Abstract* in the AO is narrower than KYOTO's *abstract* as assessed in the document 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*.
- *Quantity* – Although *quantity* is not included in KYOTO 3 Top, it is included in KYOTO 3 Middle.
- *Attribute* – This document states that KYOTO does not have *Attribute*. KYOTO 3 Top 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.*

- *Physical Attribute* – This can be compared to DOLCE's and KYOTO's *physical-region* with the similar observations for *Attribute* made above.
- *Number* – This term is not present in KYOTO 3 Top, but it is present in KYOTO 3 Middle.
- *Time Interval* – Although AO observes that their definition is more specific than KYOTO's, it is worth noting that the instances seem to be the same.
- *Space-region* - The DOLCE/KYOTO definition quoted in this document is not correct.
- *State*, *Role*, and *Disposition* – Although KYOTO does not have a class *Dependent Entity*, it is not clear from this document 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 there extensions are the same.
- *Abstract quality* – note the mistake in the definition "endurant" rather than "perdurant".
- *Collection* – The DOLCE DNS definition here is not correct. At any rate KYOTO does not consider collections to be containers. Also, notice that some instances of AO *collection* are agentive, and so do not match the KYOTO definition, e.g., a team of doctors.
- *Social-agent* – This term is not the name of a class in KYOTO, so it is not clear which class is being referred to here in the comparison.
- *Organization* – The definition cited for DOLCE-DNS, DOLCE, and KYOTO is not correct.
- *Natural person* – The definition cited for The definition cited for DOLCE-DNS and KYOTO is not correct.
- *Physical Object* – It should be noted that the class in DOLCE, and hence KYOTO, is called *physical-endurant*. The parenthetical remark in the quoted definition is not in the original and makes a difference to the meaning.
- *Organism* – While this class does not exist in KYOTO Top, it does exist in KYOTO middle.
- *Material* – While DOLCE and KYOTO do not have a class called *Material*, this class is very similar to *amount-of-matter*.
- *Artifact* – The term *material artifact* in KYOTO is inherited from DOLCE-DNS.

Until this stage, the achieved mappings between KYOTO and AO will be reevaluated based on the above review. These mappings will be used to extend KYOTO to Arabic, where each concept in KYOTO top Ontology will be mapped to its Arabic equivalent.

Framework for Mapping Between WordNet and Arabic Ontology

In this deliverable, we tried to layout the basis for one of the research activities that have to be set up within the SIERA project: the problem of creating (in a semi-automatic way) an Arabic Ontology mapped to a widely adopted lexical ontology such as WordNet. Also because the ontology creation process uses mappings among concepts in the Arabic Ontology and concepts in WordNet, one of the research areas more relevant to the ontology creation problem is Cross-Language Ontology Matching (CLOM). In the deliverable we define our problem, providing a broad definition of ontology that considers both lexical and logical ontologies, and then we discuss the state-of- the-art in the CLOM area.
However, one of the main objectives of the work carried out so far is to investigate the mapping framework that will be used in CLOM. This mapping framework includes: the representation of CLOM mappings, and a formal/ theoretical interpretation of the meaning of these mappings. For both tasks we start from definitions/approaches defined in the context of mono-lingual ontology matching and we try to extend them to CLOM, by considering the concepts' lexicalization.
We have found a recent work that introduces a classification-based semantics for weighted mappings (in mono-lingual settings) [Atencia et al., 2012], and we believe this is a good candidate for defining the intended semantics of mappings represented in CLOM. The idea behind this approach is the following: a weighted mapping between a concept C and a concept D represents the probability that an object (instance) can be classified under D if it is also classified under C (Atencia et al. interprets "classification" as "membership of an instance to a concept", according to a logical perspective).

We argue that a framework inspired by this approach [Atencia et al., 2012] solves our problem and can provide a good foundation for the notion of "semantonym" (which is defined later in the deliverable) for several reasons. Many ontology matching methods (in particular in Cross-Language Ontology Matching) use metrics that evaluate the overlap between the entities (e.g., ontology instances, documents, pieces of text) that are "classified" under two concepts. Also, the 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 (also of methods using statistical evidences).
However, a lot of work is still needed to effectively use this approach in the field of CLOM, and with respect to the kinds of ontologies we want to consider. Details about the future research directions are discussed. As an example, we have to consider the lexicalization of concepts in the definition of a mapping, as linguistic aspects are crucial in CLOM (and when considering ontologies such as WordNet and the Arabic Ontology). Since we use a broader notion of ontology, which encompasses lexical and logical ontologies, we may need to establish CLOM mappings between a logical and a lexical ontology; in this case, we need to extend the approach of Atencia and colleagues to consider different types of classified objects and different interpretations of "classification" (e.g., a term, classified under [disambiguated as] a given sense). Also, we must consider the role of communities of speakers and the type of ontologies we are dealing with, which are generic ontologies and not domain specific knowledge bases. Finally, problems such as transitivity of the mappings, have to be framed within the classification-based semantics approach.

Next we **first** identifies the heterogeneity problem and the several levels at which heterogeneity occurs. **Then** a definition of the ontology matching problem is given. In particular, it presents the currently used techniques for implementing this process. These techniques are classified with the many features that can be found in ontologies; terminologies, structures, instances, and semantics. They interview with many different disciplines such as statistics, linguistic, machine learning and data analysis. The alignment itself is obtained by combining these individual techniques in order to obtain an alignment with particular features. **After that** a general framework for cross-lingual ontology matching is introduced. In particular, we **initially** provide formal definitions for ontologies that are often involved in this process, and also precisely define the structure and the semantics of mono-lingual ontology matching. **Then** we provide an overview of the cross-lingual and multi-lingual ontology matching in relation to the mono-lingual ontology matching definition, while taking into consideration the lexicalization of the ontology entities.
Finally we propose a foundation for a cross-lingual ontology matching framework. Practically, we define what is the semantics of a correspondence for cross-lingual ontology matching, taking into account the lexicalization in the definition of the mapping.

I. INTRODUCTION

I.1. Semantic Heterogeneity

With the rapid growth of the data on the Web, there is increasing interest not only in sharing more data, but also in sharing the semantics behind this data. The notion of a *Semantic Web* was proposed [Berners-Lee et al., 2001] to deal with this massive growing amount of information and for machine understandable Web resources; to realize this, systems in the Web should be able to exchange information and services among each other semantically. Thus, the semantics of one system should be exposed (in the Web) in such a way that other systems can understand it correctly and utilize it to achieve interoperability. Various ways have been introduced and proposed in order to express, expose and understand the semantics of the various systems. This variety has lead to so-called *semantic heterogeneity*.

Ontologies have received great attention in research, as well as in industry, for enabling knowledge representation and sharing. An ontology is a structure representation of critical knowledge that enables different systems sharing this knowledge to communicate meaningfully. Ontologies are considered an appropriate answer to the problem of semantic heterogeneity.

Although the use of ontologies may facilitate semantic interoperability, multiple users or organizations are likely to declare their own knowledge ontology (domain ontology) for describing and annotating their shared documents. Accordingly, many domain-ontologies describing the same domain coexist independently created by different users. The proliferation of various domain ontologies has introduced more semantic heterogeneity. Several reasons explains this heterogeneity [Euzenat and Shvaiko, 2007] [Bouquet et al., 2004]:

1. ***terminological heterogeneity***: this kind of heterogeneity arises when different terms are used to describe exactly the same concepts, for instance when using multiple languages.
2. ***semiotic heterogeneity***: concerns the way by which entities are interpreted by people, depending on the context of usage. This is not easily distinguished by machines, because of the poor knowledge of the real context of users.
3. ***difference in coverage*****:** the domains covered are not the same, so the classes (concepts) used in the ontologies do not represent the same things in two different ways.
4. ***difference in granularity*****:** it occurs when the same domain is described by different levels of details.
5. ***difference in perspective***: the point of view of an ontology engineer lets him emphasize some particularity leading to a specific hierarchy for concepts based on his own criteria. This may result in multiple ways of creating an ontology and leads to a strong heterogeneity.

As a result, new form of heterogeneity have been introduced, that is, the *ontology heterogeneity*. Now, current approaches mostly tackle the problem by *matching ontologies*, that is, by finding correspondences between semantically related ontological entities[Bouquet et al., 2004]. This field is very active and has attracted a lot of attention in the last few years, but is far from being resolved [Euzenat and Shvaiko, 2013].

I.2. Ontology Matching Overview

In response to the generalized heterogeneity on the growing amount of published ontologies on the Web, in the last two decades a specific research field has emerged, the so-called *Ontology Matching*. Ontology matching studies the ways to automatically establish semantic relationships (correspondences) between two (or more) ontology entities [Euzenat and Shvaiko,2007].

Ontology matching enables ontologies to interoperate. However, discovering (automatically, or even manually) such correspondences between different ontologies is a complex task, deep reasons of heterogeneity between ontologies to be matched are not explicitly known by machines (and for human to some extent) as explained before.

In general, matching methods are combinations of individual (atomic) matching techniques [Shvaiko, 2004] that can be divided into four categories based on which kind of data the matching techniques work on [Shvaiko and Euzenat, 2013]: terminological techniques, structural techniques, instance-based(extensional) techniques, and logical reasoning(semantic) techniques.

The *terminological techniques* (or, in general element-level techniques [Euzenat and Shvaiko,2007]) refers to the *string-based* and *linguistic-based* techniques that find correspondences between the ontologies textual entities descriptions and labels. String-based metrics take advantage of similar characters from two strings; whereas

linguistic-based metrics compare the meaning of strings. The underlying idea is that the more similar two entities' strings are, the more they are likely referring to the same concept. Various string-based techniques have been proposed to compute this similarity; one can simply compute the longest size of common substrings (prefixes, suffixes), or more sophisticated ones such as edit distance (e.g. Levenstein distance, Monger-Elkan distance, Jaro-Winkler distance), different string matching algorithms can be used here; for more details on string similarity methods refer to [Cohen et al.,2003]. Another technique leverages on linguistic tools as a pre-processing phase before the string-based comparison, making use of various NLP techniques (e.g., tokenization, lemmatization and stemming) in order to exploit their morphological properties. Linguistic resources (like common knowledge, domain specific thesauri, linguistic ontologies, or dictionaries) also introduced to bridge the gap between a syntactic piece of information and its meanings. For instance, WordNet [Fellbaum,1998] gives all the senses of a given word (called synsets), and provides a directed relation graph between the synsets that represents the semantic relations between synsets. Comprehensive details on using WordNet for ontology matching can be consulted in [Lin and Sandkuhl,2008].

The *structural-level techniques* [Euzenat and Shvaiko,2007] make use of the structural presentation of ontologies. The structural-based approaches consider the ontology as a graph whose nodes represent the ontological entities, and the edges are labeled with relation names. The problem of matching ontologies is viewed as a problem of matching graphs. The underlying assumption of it is based on the fact that the similarity between two entities on two respective graphs impacts the similarities between the respective neighbor entities in each graph, this idea can be grounded in several ways; by comparing the nodes (entities) children, leaves, or comparing entities in the transitive closure, among others.

The basic idea of *instance-based* (extensional-based) mapping techniques is based on the analysis of statistics or distributions of class extensions; the more common instances of two concepts are, the more they are likely to denote the same concept [Isaac et al, ISWC2007]. Instance-based techniques can also rely on instance properties or descriptions. Instance analysis can be exploited to compute similarity scores between classes or to train classifiers for machine learning methods.

The *logical reasoning* (m*odel, or semantic*) based techniques exploit the semantic interpretation of the input ontologies and apply methods like propositional satisfiability (SAT) or description logics reasoning techniques to check the logical consistency of the candidate correspondences returned by previous steps, or to deduce other correspondences from the previously discovered ones.

[Mochol, 2009] has proposed a deeper classification of the matching systems, taking into account several dimensions. These dimensions are:

1. *Input characteristic*, takes into account the type of ontologies to be matched; depending on their size, expressiveness and formality (glossary, thesaurus, taxonomy, schema, and ontology), language and role, e.g. domain ontology or upper-level ontology. Also, the use of external resources (e.g., matching rules, domain constraints, dictionary, or previous matching decisions).
2. *Approach characteristic*, describes the matching algorithms; the matcher type (individual, or combination), the input interpretation of the matcher as an element, structural, instance, or semantic based method, the way of processing the algorithm (manual, semiautomatic (user intervention), or automatic), among other futures.
3. *Output characteristic* defines the desired result of the matching execution, considers the output type (e.g., relation, similarity measure), the matching cardinality (1:1, ?:?), the execution completeness (e.g., partial, full).
4. *Usage characteristic* considers the different situations in which the approaches can be or have already been used, for instance, the goal of usage , and the application area, and adaptation ability to be used in different domains and applications.
5. *Documentation characteristic* points out the existence and quality of the documentation.
6. *Cost characteristic* addresses the costs for the usage of an algorithm.

More on classifications of matching methods can be found in [Bouquet et al. 2003, Doan and Halevy 2005, Ehrig 2007, Euzenat and Shvaiko 2007].

II. STATE-OF-THE-ART

As outlined in the introduction, *Ontology Matching* is a solution to the semantic heterogeneity problem. It establishes semantic relationships among ontological resources between independent ontologies. The last two decades have witnessed a wide range of ontology matching methods, which have been successfully developed and evaluated [OAEI, 2005-12]. Several recent surveys [Choi et al.,2006, Shvaiko&Euzenat,2013] and books [Euzenat&Shvaiko,2007, Bellahsene et al,2011] have been written on this field, also several conferences and workshops[OM-12] have specifically tackled this topic as well.
The majority of the proposed matching techniques in these systems have mainly focused on mapping between ontologies that are lexicalized in the same natural language (so-called, *Mono-lingual Ontology Matching*, MOM). Moreover, methods developed for MOM systems cannot directly access semantic information when ontologies are expressed in different natural languages. However, there is a need for a method that automatically reconciles information when ontologies are lexicalized in different natural languages [Gracia et al.,2012].
Recently, notable efforts [Spohr et al.2011, Fu et al.2012, Trojahn et al.2010] 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-Lingual 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 the context of mapping the Arabic Ontology to WordNet next we focus on the cross-lingual ontology matching case. We start by giving an overview of recent efforts in this domain. In particular, we shall focus on methods and techniques for cross-lingual sense disambiguation. Finally, we conclude with an analytical comparison between these cross-lingual matching approaches.

II.1. Cross-Lingual Ontology Matching

In this section we will specifically study the case of cross lingual ontology matching techniques. In general, to resolve the cross-lingual issue, a translation based approach is considered in order to transform the cross-lingual problem into a mono-lingual ontology matching one. The translation-based approach leverages on machine readable dictionaries (mainly, bi-lingual dictionaries) [Nagi et al.2002, Liang et al, 2006], and machine translation tools (e.g. Google, Bing,.. etc.) were also used to translate the ontological resources [Spohr et al.2011, Fu et al.2012, Trojahn et al.2010].

[Liang et al, 2006] used a bilingual *dictionary* (Chinese-English) to overcome the language barrier, whereby mappings are generated manually (by human experts). The English thesaurus: AGROVOC (developed by the FAO containing a set of agricultural vocabularies) is mapped to a Chinese thesaurus: CAT (Chinese Agricultural Ontology, developed by the Chinese Academy of Agricultural Science). The thesauri are loaded in the Protégé editor, and segments of the thesauri are assigned to groups of terminologists to generate mappings. The mappings generated by such approaches are likely to be accurate and reliable. However, this can be a resource consuming process specially for maintaining large and complex ontologies.

An *unsupervised* method was suggested based on (non-parallel) bilingual corpora [Nagi et al.,2002]. Nagi et al. also used a bilingual dictionary (Chinese-English) in mapping between HowNet (in Chinese) thesaurus and WordNet (in English). Nagi et al. rely on the available synsets in both resources to find the proper mapping. This approach, as it happens with most unsupervised learning methods, heavily relies on corpus statistics. In some cases, highly frequent patterns with incomplete semantic meaning may be produced, apart from the corpora construction overhead.
A pseudo feedback was adopted by Fu et al. in order to improve the matching quality by assessing the generated matches if they are above a certain threshold (without user intervention) [Fu et al.2012]. Fu et al. approach, alongside the machine translation process, exploits the structural information of the ontologies by considering the context of the entities to be matched, which is defined by the set of neighboring (immediate surrounding) entities.

As an alternative, *machine learning* techniques were introduced to solve cross-lingual issues, for instance, Spohr et al. used a Support Vector Machine (SVM) to learn a matching function for ontologies represented in different

languages [Spohr et al.2011]. At first they translated the ontology labels to a pivot language (English) via a machine translator, then with a combination of string-based and structural-based similarity metrics they built the feature vector. This system, deeply leverages the structural information derived from the ontology. Furthermore, this approach, like all supervised learning methods, requires a significant number of labeled training samples and well designed features to achieve good performance.

The *Instance*-based matching techniques were also exploited in this direction. Such an example is presented in [Wang et al. 2009]. Wang et al. use a machine translation service to translate a digital library vocabularies written in English, French and German, then map the concepts occurring in subject heading lists, which are often used to describe objects from library collections. Wang et al. determine the similarity between concepts by examining the overlapping of similar instances classified with the concepts.

Another interesting work for resolving the cross-lingual issue exploits Wikipedia for external knowledge. For instance, [Hertling and Paulheim et al. 2012] search Wikipedia articles (pages) title for a given term (labels, and comments) and retrieve all language links describing the term, making use of the inter-language links between Wikipedia pages. Then they compare the retrieved titles with the same language and return the maximum of the cross product from label and comment, by computing the Jaccard coefficient of the two sets of retrieved titles.

In spite of these notable efforts, introduced above, [Spohr et al.2011] argued that the quality of machine translation systems is limited and depends greatly on the pair of languages considered. Moreover, translation tools (to some extent) might remove the language barrier but not necessarily the cultural one; there is the need to find the appropriate (*sense*) concept of the translated word and not only to laterally translate them [Cimiano et al. 2010].
An interesting approach is to disambiguate and discover the proper semantic (sense) of keywords, more than just exploit machine or dictionary translations. In particular, the work of [Trillo et al. 2007, Melo&Weikum 2008, Navilgi&Ponzetto 2010], is discussed next.

II.2. Disambiguation and Sense Selection

An interesting approach for disambiguating and liking cross-lingual senses was proposed in BabelNet [Navigli and Ponzetto, 2010]. BabelNet is a large multilingual semantic network built from Wikipedia and WordNet, which provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.
In particular, this is done by assigning WordNet synsets to Wikipedia entries (pages titles). Since Wikipedia entries allow inter-lingual relations, a variant in several languages can be assigned to some of the WordNet synsets. Therefore, if a relation between a synset and an English Wikipedia entry has been set, using the Wikipedia inter-lingual links the same relation can be set for all languages having the corresponding Wikipedia entry (in other languages). The mapping between Wikipedia pages and WordNet senses was made based on a probabilistic estimator that exploits the structural information available in WordNet (e.g., synonyms, hypernymy/hyponymy, and gloss) and Wikipedia (e.g., redirect pages, disambiguation pages, internal links, inter-lingual links, and categories) in order to build the so-called context of WordNet sense and context of Wikipedia page. At the end of this step, the initial Babel Synset was created.
However, Wikipedia provides incomplete socio-cultural knowledge for different languages as inter-linguagl links do not exist for all articles. For those languages lacking the corresponding Wikipedia entry, [Navigli and Ponzetto, 2010] proposed the use of a machine translator to automatically translate a set of English sentences containing the synsets (Babel Synsets), this set of sentences is built using two sources: SemCor [Miller et al., 1993], a corpus of more than 200,000 words annotated with WordNet senses, and sentences from Wikipedia containing a link to the English Wikipedia page. After they applied the automatic translation, the most frequent translation for each specific term (Babel Synset) is detected and included as a variant for the Babel synsets in the given language.

Trillo et al. 2007 proposed a system to disambiguate user keywords in order to translate them into semantic queries. In this context, a semantic similarity measure has been defined to provide a synonymy degree between two terms from different ontologies, by exploring both their lexical and structural context. In the end, configurable thresholds allows the system to determine whether to consider the two ontological terms as similar or not.

Specifically, the Trillo et al. approach discovers the possible meanings of a set of user keywords by consulting a pool of online available ontologies (accessed by Watson[9] or Swoogle[10]), also other local ontologies and lexical resources, as WordNet [Fellbaum, 1998]. Then it proposes a set of possible ontological senses for each keyword, then based on synonym probability measurements iteratively integrating the ones that are considered to be similar enough, in order to reduce any redundancy. In particular, each ontological term is compared with the rest of ontological terms, and its similarity degrees are computed by using synonym probability measures. If the similarity degree between two terms is greater than a threshold given as a parameter (synonym threshold), then they are considered synonyms (i.e., they represent the same interpretation of the keyword) and they are merged (integrated) into a single sense following the techniques described in [Trillo et al. 2007]; otherwise the two terms are considered to represent different meanings of the keyword. The output of the process is a set of (single or integrated) senses, where each element corresponds to a possible meaning of the keyword. At the end, a disambiguation process is run to pick up the most probable sense for each keyword, according to the context, to use them in the construction of semantic queries.

A machine learning approach was proposed by [Melo and Weikum 2008] to determine the appropriate senses among the translated candidates. The proposed approaches where developed to automatically produce a WordNet like resources for new languages. In conjunction with the English WordNet, they relied on the Ding German-English dictionary [Richter, 2007] which does not always have a part of speech tags. At first, a set of candidate mappings are achieved with a direct translation, then the associations between terms and senses ate predicted. To determine the appropriate senses among the candidates they constructed a binary classification learning problem. To create the feature vectors, several scores that take into account structural properties as well as semantic relatedness and corpus frequency information where used. Comparing the overall results achieved with the literature work in automatically constructing (WordNet like) lexical resources, the authors claimed that, this technique is imperfect in terms of their quality and coverage of language-specific phenomena. Nonetheless, it can be useful for alternative applications e.g., thesaurus generation, semantic relatedness, and cross-lingual information retrieval.

II.3. Analytical Comparison

An overview of recent cross-lingual matching systems is presented in Table 1. The first column provides a general overview of the system. The mediator column expresses the use of external knowledge that exploits the semantics of labels and are based on tools that explicitly codify semantic information (e.g. thesauruses, dictionaries and machine translations, other ontologies). The input column presents the type of source and target ontologies used by the matcher; thesaurus, lexical-based or axiomatic-based ontologies. The user interaction column expresses the level of automation (manual, (semi)automatic).
The table also classifies the available matching methods, depending on which kind of data the algorithms consume; (*Terminological*) linguistic-based and string-based techniques find correspondences between textual entities, descriptions and labels. String-based techniques match entity labels or descriptions syntactically. The underlying idea for matching entities with their names is that the more similar the names are(according to a chosen measure), the more there are likely to denote the same concepts. (*Structural*) Structure-based techniques exploit the ontology internal structure, as well as the relations among entities. (*Extensional*) Instance-based matching techniques are based on the analysis of statistics or distributions of class extensions. (*Semantics*) A model which gives a semantic interpretation usually uses the reasoning process to infer correspondences from previously discovered correspondences. The last column presents the evaluation methodology.

[9] Watson: http://kmi-web05.open.ac.uk/WatsonWUI/
[10] Swoogle : http://swoogle.umbc.edu/

Table 1: Analytical comparison of CLOM

System	Overview	(Mediator) *Background/External knowledge*	Input Type	User interaction *Automation level*	Matching method				Evaluation-based
					Terminological	Structural	Extensional	Semantic	
BabelNet [Navigli and Ponzetto, 2010]	Integrate and Map WordNet to Wikipedia based on a probabilistic mapping, using machine translation, and Wikipedia inter-link.	SemCor, machine translation, and Wikipedia inter-link	lexical ontology: WordNet. Online collaborative resources: semi-structure information (Wikipedia)	automatic	WikiPage titles, WordNet senses,	Wiki-, and WordNet- Context Mapping probability: Bag of word(context overlapping), or Graph based (graph connectivity)			Gold standard; 1k WikiPage mapped to WN, and automatic mapping vs. EuroWN
[Melo and Weikum, 2008]	Machine learning approach to build aligned wordnet. Using bilingual dictionary for translation, and a Binary classification (SVM, feature vector: incorporate different scores)	Bi-lingual dictionary	lexical ontology: WordNet.	Needs a training set	Wordnet, gloss, cosine similarity, TF.IDF,	WordNet structural (dijkstra-like algorithm)			Gold stander and vs. other methods

[Trillo et al. 2007]	Sense disambiguation using machine translation, and external knowledge: Pool of ontologies	Pool of ontologies	ontologies		Synonyms, Jaro-Winkler metric, Normalized Google distance	hypernym/hyponym relationships.		Iterative sense alignment Synonymy probability between two keyword senses	
UWN [Melo and Weikum, 2009]	Translation based for Building a universal wordnet Based on [Melo and Weikum, 2008]	Wikitionary, machine translation, mono/multi-thesaurus, parallel corpora's and manually			Bootstrapped by EuroWordNet	Iterative graph based, statistical learning; SVM,			Gold stander to evaluate the learning
SOCOM++ [Fu et al.2012]	Configurable system, based on pseudo feedback(confidence based > user threshold), and machine translation.	machine translation			Yes	Yes			2 gold standard
[Spohr et al. 2011]	Machine learning, using machine translation, and similarity aggregation	Used AROMA (2006) ontology alignment			Levenshtien, BOW-cosine, substring distance,	Calculation information(direct and elementary children)		Ranking SVM over both string, and structural features	reuse (multilingual Ontologies) as gold standard
[Trojahn et al. 2010].	Direct and indirect translation based alignment, using machine translation, and combined different string-based methods	WordNet, and machine translation	DL Ontologies		Synonyms, and string-based methods: substring, edit distance, name equality matching, string commonalities (all				OAEI-08,09

					substring match) and differences(all un-match substring)				
Thesaurus Mapping: (En) AGROVOC and (Ch) CAT [Liang and Sini, 2006]	Manually based mapping, using a (Ch-En) bilingual dictionary	(Ch-En) bilingual dictionary	Thesaurus		Stemming, String Matching	Thesaurus term relations			Expert review
Thesaurus Alignment: (En) WordNet and (Ch) HowNet [Ngai et al., 2002]	Corpus based mapping based on term frequency analysis (vector-based co-occurrence),and synset mapping. using a (non-parallel) bilingual corpora, and (Ch-En) bilingual dictionary	(non-parallel) bilingual corpora, and (Ch-En) bilingual dictionary	Thesaurus, and lexical ontologies(Wordnet)		direct hypernym/hyponym-set				160 definition, expert annotation
Thesaurus Mapping Dutch (GTAA, GTT) to (En) WordNet [Malaise et al,2007]	Thesaurus mapping to wordnet using the lexical description (gloss) overlapping, and a bilingual dictionary	a bilingual dictionary (supports definition and POS)	Thesaurus, and linguistic ontologies(Wordnet)		Stemmers, and Gloss overlapping				Manual test sample
RiMOM [Zhang et al.,2009]	Translation based mapping using a (Jp-En) bilingual	Bi-lingual dictionary(Jp-En)			Edit distance, Wordnet, Vector distance	Similarity propagation (path-similarity)	Vector distance	-	OAEI -08 mldirctory set

	dictionary								
RiMOM [Wang et al., 2010]	Multi-strategy dynamic ontology alignment. Automatically and dynamically combine multiple strategies; considering textual and structural characteristics	Wordnet. The translation based they follow is not clear.	OWL ontologies		Edit distance, Wordnet, Vector distance	Similarity propagation (Flooding algorithm)	Vector distance	-	OAEI -10 VLCR set
GG2WW [Bouma,2009]	A translation based for mapping (DE) GATT thesaurus to (En)Wordnet and (En)DBpedia (OAEI-09). Using Wikipedia and EuroWordNet inter-lingual.	EuroWordNet, Wikipedia			Stemmer, EuroWordNet, Wikipedia (En-Du) inter-lingual	EuroWordNet's (synonym, and near-synonym relations)			randomly select, and evaluated against 100 gold standard
DSSim [Nagy et al., 2009]	A translation based mapping (DE) GATT thesaurus to (En)Wordnet and (En)DBpedia (OAEI-09). Using DBpedia inter-lingual	Wordnet, DBpedia							randomly select, and evaluated against 100 gold standard

YAM++ [Ngo, and Bellahsene, 2012]	A machine learning approach (Decision Tree, SVM, NaiveBayes). for combining several metrics, which exploits various string and linguistic based metrics. Using machine translation, And an Information retrieval techniques (term frequency).	Wordnet, Machine Translation		Training set	String-based metrics: used SeconString, and SimMetric open-source libraries, and Equality, Prefix, Suffix, Longest Common SubString and Stoilois. linguistic-based metrics: WordNet similarity metrics (Lin, JiangConrath and Wu- Palmer).	Similarity propagation (Flooding algorithm)	instances belonging to the class or its descendants.	The Global Optimal Diagnosis method (global constraint optimization method proposed in Alcomox tool)	OAEI-12 Multifarm set
WikiMatch [Hertling and Paulheim, 2012]	Exploits the Wikipedia search functionality and multilingual inter-links	Wikipedia			WikiPages name overlap				OAEI-12 Multifarm set
MapSSS [Cheatham, 2012]	Combines syntactic, structural, and semantic matrices.	Google Research, Machine translation			Levenstein distance	Direct neighborhood (entities are same type, same edge label, VF2 graph matching		Google *Research* API for synonym detection	OAEI-12 Multifarm set

						algorithm)			
WeSeE [Paulheim, 2012]	Use Web search engine for retrieving web documents that relevant for concepts in the ontologies, and machine translation based.	Web search engine, and machine translation			TF-IDF measure				OAEI-12 Multifarm set
GOMMA [Gro et al., 2012]	Performs direct (internal ontology knowledge) and indirect (existing mappings to intermediate (background ontologies)matching. Iterative build bilingual dictionary using a machine translation based	Machine translation			Normalizing, name/synonym matcher	Blocking strategy (graph based)		mappings consistency	OAEI-12 Multifarm set
AUTOMSv2 [Kotis et al.,2012a] ASE [Kotis et al.,2012b]	Combines several methods; Lexical-, structural-, instance-based. using machine translation,	Wordnet, machine translation	OWL		partition-based clustering: COCLU. String-based; 'smoaDistance' method and the 'levenshtein Distance'.	Class Properties' similarity(String matching method (Levenshtein))	Class Instances' similarity (String matching method	vector-based LSA (Latent Semantic Analysis) and WordNet	OAEI-12 Multifarm set

					WordNet synonym based similarity		(Levenshtein)		
MEDLEY [Hassen, 2012]	Use a lexical metrics and structures matching between links of each node. using machine translation based,	machine translation	OWL		Lexical treatment : q-gram and levenshtein distance , tokenization, stemmer.	the node that neighbor is aligned to must be a neighbor of any prospective match			OAEI-12 Multifarm set
CODI [Huber et al., 2011]	Aggregating different similarity measures; a Probabilistic-logical alignment. Matching different versions of the same ontology using machine translation.	machine translation			Tokens, normalization, combine several string similarity		object-property assertions	syntax and semantics of Markov logic	OAEI-12 Multifarm set
LogMap [Jimenez-Ruiz et al, 2012]	logic-based ontology matching system	The translation based they follow is not clear.	Logical		Lexical indexation			Logic-based module extraction. Propositional Horn reasoning.(Dowling-Gallier algorithm)	OAEI-12 Multifarm set

[Wang et al., 2009]	Mappings among library subjects written in English, French and German Applying lexical and instance-based matching techniques. using machine translation	Machine translation	SKOS		adapts [Malaise et al, 2007]		instances overlapping , and instance matching(tf-idf weighting over the instances metadata)		Gold standard
[Lin, Feiyu et al. 2011]	Combines several methods; string, lexical, structural and context based. Using machine translation, and Wikitionary for translation	Machine translation and Wikitionary			Jaro-Winkler distance(SimMetrics) and SmithWaterman algorithm(SecondString) , WordNet, and Jiang-Conrath measure(Wordnet-based).	Ontology triple overlap, Subclass similarity, Expanding tree method			OAEI 101, 206 set

III. PRELIMINARIES

In this section we define what we mean with ontology, in particular, the different ways ontology is used in ontology matching problems, which includes both lexical and logical ontologies.
Then we provide a formal definition of the mono-lingual ontology matching problem, and define the structure and semantics of mappings in the mono-lingual ontology matching. Next we provide an overview of the cross-lingual and multi-lingual ontology matching in relation to the mono-lingual ontology matching definition, considering the ontology lexicalization.

III.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 $\boldsymbol{\mathcal{O} := (\mathcal{C}, \mathcal{E}_R, \mathcal{T}, \mathcal{I}, \mathcal{R}, A)}$
where,
$\boldsymbol{\mathcal{C}}$ is a set of classes (or concepts). $\boldsymbol{\mathcal{E}_R}$ *is* a set of relations between classes (e.g., hyponymy($\preccurlyeq$), equivalence ($\equiv$), subsumption ($\sqsubseteq$), or disjoint($\perp$)). $\boldsymbol{\mathcal{T}}$ is the set of all possible ontology entitiy labels (concepts, relations, comments,..etc) 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 $c \in \mathcal{C}$. $\boldsymbol{\mathcal{R}} \subseteq \mathcal{I} \times \mathcal{I}$ is a set of relationships between instances, and $\boldsymbol{A}$ is a set of axioms in a logical language on $\boldsymbol{\mathcal{O}}$.

Definition 1 provides a broad definition that encompasses the use of terms for referring to several knowledge representation systems. However, as introduced in the state of the art section, matching systems consider different types of ontologies to be matched [Mochol, 2009]. Thus, we can specialize our definition to reflect deferent type of ontologies. For instance; the deferent ways that ontologies are used in ontology matching problems encompasses both lexical and logical ontologies.

- ***Lexical Ontology**:*
  Lexical ontology (or linguistic ontologies) can be defined as $\boldsymbol{\mathcal{O}^{Lex}} = \{ \boldsymbol{\mathcal{C}, \mathcal{E}_R, \mathcal{T}} \}$, where: $\boldsymbol{\mathcal{C}}$ represent a set of *synsets*. $\boldsymbol{\mathcal{E}_R} \subseteq \mathcal{C} \times \mathcal{C}$ is a set of relations, both lexical (e.g., synonym, antonym) and semantic (e.g., sup/super-type of). $\boldsymbol{\mathcal{T}}$ is the set of all possible synsets lexicon(terms) in a given natural language, represented as the set $\mathcal{T} = \{t_i, \dots t_n\}$.

- ***Logical Ontology:***
  Logical ontology can be defined as $\boldsymbol{\mathcal{O}^{Log}} = \{ \boldsymbol{\mathcal{C}, \mathcal{E}_R, \mathcal{T}, \mathcal{I}, \mathcal{R}, A} \}$, where: $\mathcal{C}$ is a set of *concepts*. $\boldsymbol{\mathcal{E}_R} \subseteq \mathcal{C} \times \mathcal{C}$ is a set of relations. $\boldsymbol{\mathcal{T}}$ is the set of all possible ontology entity labels (concepts, relations, comments,..etc) in a given language, represented as $\mathcal{T} = \{t_i, \dots t_n\}$ where $t_i \in \ell_j$ such that $\ell_j \in \mathcal{L}$. $\boldsymbol{\mathcal{I}}$ is all the possible sets of objects (instances) classified under the a class $c_i \in \mathcal{C}$. $\boldsymbol{\mathcal{R}} \subseteq \mathcal{I} \times \mathcal{I}$ is a set of relationships between objects that are members of concepts, and $\boldsymbol{A}$ is a set of axioms that define constraints on the domain, where the relation sets $\boldsymbol{\mathcal{E}_R}$ and $\boldsymbol{\mathcal{R}}$ are defined.

In this document we use the notion *Ontology* to convey all the mentioned types, otherwise we specifically differentiate the lexical or logical type.

III.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 a 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 ontological entities($\mathcal{C}$) 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 can be either a manual, or a (semi)-automatic method, 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[11]. 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 Figure1.

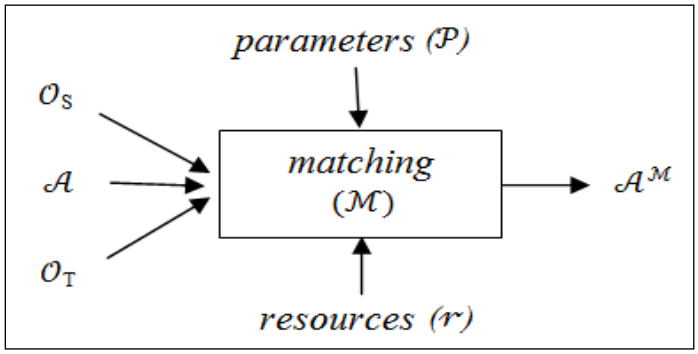

Figure1: The ontology Matching process [Euzenat and Shvaiko, 2007].

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

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

Definition 4: (Correspondence). [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 $r \in \mathcal{R}$, 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).

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;

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

III.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 ***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, $\boldsymbol{\mathcal{T}}^{\ell,\mathcal{O}}$, if the ontology terms $\boldsymbol{\mathcal{T}}$ are ***lingualized*** 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 $\mathcal{O}_T$, which lexicalized in a set of natural languages $\mathcal{L}_S$ and $\mathcal{L}_T$, respectively, and $\boldsymbol{\mathcal{T}}^{\mathcal{L}_S,\mathcal{O}_S}$ and $\boldsymbol{\mathcal{T}}^{\mathcal{L}_T,\mathcal{O}_T}$ be the set of terms (labels) of $\mathcal{O}_S$ and $\mathcal{O}_T$ lingualized 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 relationships or 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 $\boldsymbol{\mathcal{T}}^{\mathcal{L}_S,\mathcal{O}_S}$ and $\boldsymbol{\mathcal{T}}^{\mathcal{L}_T,\mathcal{O}_T}$ 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 relationships or 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 $\boldsymbol{\mathcal{T}}^{\mathcal{L}_S,\mathcal{O}_S}$ and $\boldsymbol{\mathcal{T}}^{\mathcal{L}_T,\mathcal{O}_T}$ in at least *two* languages, with $|\mathcal{L}_S \cap \mathcal{L}_T| \geq 2$.

Definition 8: ***Cross-Lingual Ontology Matching (CLOM),***
CLOM is the process of establishing relationships or 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 $\mathcal{T}^{\mathcal{L}_T,\mathcal{O}_T} \in \ell''$

IV. PROBLEM STATEMENT

In the context of cross-lingual mapping, the language 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,Fu et al. 2009,Wang et al. 2009].
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 semantic mappings that align concepts lexicalized in different natural languages. In general, a community of users (speakers) would consider two concepts that are differently lexicalized in a given language to be equivalent if both lexicons are used to indicate the same meaning in a given *context.*
The context (or *context of discourse*) that a community (of language speakers) shares in order to decide if these two terms (lexicon) refer to the same concept is "*not only explain what people say, but also how they say it. Lexical choice, syntax, and many other properties of the 'formal' style of this speech are controlled by the parliamentary context*" [Van Dijk,2006].
Moreover, [Lichao Song,2010] argues that the context in which human beings explain what is in their minds depends on three factors; (Linguistic) the relationship between the words, phrases, sentences and even paragraphs. (Situation) the environment, time and place, etc. in which the discourse occurs, the relationship between the participants. (Cultural) the culture, customs and background of period in language communities in which the speakers participate. He also reports that languages are a social phenomenon that strongly tied to the social structure and value system of society.

Accordingly, mapping between concepts that are lexicalized in different languages, indeed is a challenging task, as said before not only because of the language barrier but also because of the cultural one. Without loss of generality, if two concepts are lexicalized in different languages, then they are considered equivalent if they express and indicate the same meaning in a given context. That is, if both language communities (the majority of language speakers) share the same context (interpretation) for a given concept, whatever the lexical notation being used (language), as they refer to the objects (entities) belonging to this concept.

The goal is to provide a formal definition of the cross-lingual ontology matching problem, mainly to define what a correspondence is, and how to represent correspondences in cross-lingual 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 want to find a good foundational interpretation of the semantics of the correspondence in the CLOM, we do this by selecting a theory from the state of art that we believe is helpful for defining what is the intended semantics of the correspondences in CLOM in the classification based semantic we identified a good candidate.
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 this semantic on the use of the concepts and the classes for objects in a logical domain, 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 based on the idea of using sets of information objects classified with a concept in order to submit the translation.
However, at the same time, the adaptation of the framework to CLOM is not trivial. In the next sections we first introduce an approach proposed in the literature for MOM for logically founded ontologies, in particular the approach presented in [Atacia et al. 2012]. Then we extend to CLOM for logical ontologies.
However, the Atacia et al. approach does not explicitly introduce the lexicalization of the ontologies, which are very important and fundamental aspects used by ontology matchers [Euzenat and Shvaiko 2007]. As a result, we need to provide as a first step a lexicalized version of the ontology matching problem mainly in the context of CLOM for logically founded ontologies based on classification interpretation of mappings. Then we explain how to extend this approach to also include the matching of ontologies that are not straight forwardly interpreted as logical ontologies or can be mixed as logical and lexical ones.

V. Classification-based Interpretation of Mappings in Mono-Lingual Logical Ontologies

Ontology mapping which can be seen as an expression that establishes relations between elements of two (or more) heterogeneous ontologies, a crisp mapping tell us that a certain concept related to another concept in different ontologies, and the type of relations are typically set of theoretical relations { $\equiv, \sqsubseteq,$ or $\perp$ }, while the weighted mapping in addition associates a number (weight) to those relations.
An interesting approach presented in [Atencia et al. 2012] provides 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 on a 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 $\{x_1, \ldots, x_{10}\}$, 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 $\{x_1, \ldots, x_{10}\}$, classified under the concepts $C \in \mathcal{O}_1$ and $D \in \mathcal{O}_2$, then we say that concept C and D are euivelant 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 (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 a two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, a weighted mapping from $\mathcal{O}_1$ to $\mathcal{O}_2$ is a quadruple:

$$\textit{Weighed Mapping:} = < C; D; r; [a,b] >$$

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].

V.1. Classificational Semantics For Weighted Mappings [Atencia et al. 2012]

V.1.1.Logical Semantics

Following the standard model-theoretic logical 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 $C^{\mathfrak{T}} \subseteq \Delta^{\mathfrak{T}}$ for $C \in \mathcal{C}$, $r^{\mathfrak{T}} \subseteq \Delta^{\mathfrak{T}} \times \Delta^{\mathfrak{T}}$ 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, . . . , 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."*

V.1.2.Weighted Mapping Semantics

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 $C_X^{\mathfrak{T}_1} = \{x \in X \mid x^{\mathfrak{T}_1} \in C^{\mathfrak{T}_1}\}$ 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 to define 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 a pointwise 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 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*.

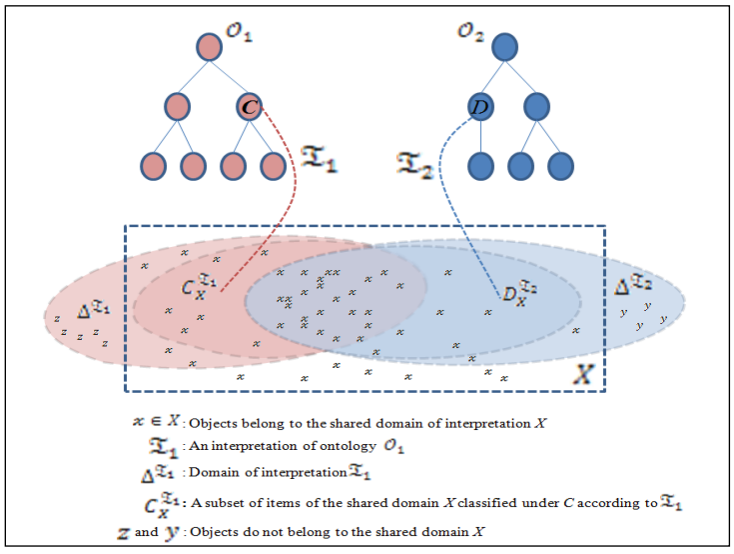

Figure 2: The extensional meaning of a concepts

V.1.3. Challenges and Open Issues

In the original definition of the extensional meaning of a mapping using the classification based approach that Atencia et al. proposed assumes a logical interpretation as a concept denoted as class of instances in an interpretation domain. The extensional meaning of a concept is interpreted as a subset of objects in a shared domain of interpretation provided by $(\Delta^{\mathfrak{T}_1} \cup \Delta^{\mathfrak{T}_2} \cup X)$.
In the logical domain the interpretation of classification is a concept that classifies individuals(objects), where these individuals are members of a class. The extensional meaning in this case cannot be directly adapted to ontologies that do not have such a logical interpretation of classification. For instance, when we annotate a document we can consider the concept as classifying an object, but the interpretation of classification is different; in this case saying that a concept classifies an object means that the concept is the topic of the document. While if we consider a text where we have several terms and we want to provide a disambiguation of the meaning of the term, we can classify a term with a concept saying that the sense of the term is the associated concept.
We claim that one can extend the extensional interpretation of mapping in the logical domain for other types of ontologies using different ways of interpretation of extension and different interpretation of the notion of classification of an instances with a concept. Besides that the Atencia et al. classification approach considers a finite set of objects belonging to the shared context of interpretation, while if we consider generic ontologies representing concepts lexicalized in languages that are spoken by a very large community, the shared context (or domain) of interpretation of the mapping problem might be very large or even infinite because some concepts refer to objects that might have an infinite extension. The question here is what is the impact on this formalism if we consider an infinite set of objects in the shared context.
Moreover, the proposed approach represents a semantic mapping between two ontologies belonging to the same type of interpretation, logical ontologies in this case. An interested research direction might be the study of mapping two ontologies, which are interpreted in a different way (cross-ontology interpretation), i.e., can such semantics be extend to map mixed interpretations, e.g., lexical and logical ontologies.
We argue that such an approach can fit the CLOM problem. However, in ordered for us to adopt our border notion of ontology, which encompasses lexical ontologies and logical ontologies, the classification based mapping approach presented in [Atencia et al. 2012]. We need to extend this definition using the classification based approach *independently* of the interpretation of classification and the type of objects that can be classified under the concept, as well as 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 cross-lingual matching problem.

VI. Classification-based Interpretation of Mappings in Cross-Lingual Logical Ontologies

The extension of concept is often used in many cross-lingual ontology matching strategies; this extension is interpreted in different ways, e.g., instances classified under concepts, objects occur in a concept, 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 based on the classification of different kind of objects with a concept to interpret the semantics of mapping.
First we extend the notion of a mono-lingual matching definition to the cross-lingual matching one by considering the lexicalization of the ontology entities in a logical domain. Then we elaborate on a classification based interpretation of mappings to define the semantics of mappings; we use 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. Finally, we extend the Atencia et al. definition by providing a lexicalized version of the classification based interpretation of the meaning for weighted mappings in a logical domain.

Let $X = \{x_1, \dots, x_n\}$ 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.

The *F-measure* of $C_{\mathcal{I},\ell}^{\mathfrak{T}_1}$ and $D_{\mathcal{I},\ell'}^{\mathfrak{T}_2}$ 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 pointwise interval can be used; that is, we assume that a=b, then n=[a, a]. Thus, we can simply present the mapping relation as $\langle C_\ell, D_{\ell'}, r, n\rangle$.

VI.1. The Semantonym Mapping

We introduce our notion for semantic mapping between two concepts lexicalized in different languages, called ***semantonym***, a cross-lingual mapping based on a classification based approach.

Intuitively, two concepts lexicalized in different languages are considered to be *semantonym*. 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 certain trust degree (confidence level) *n*. Various approaches can be adopted here to measure the confidence level based on the interpretation of the extensional model as well as the type of the ontologies to be matched and their level of formalism (e.g., lexical or logical ontologies). For instance, a probabilistic one can be adapted similar to [Atencia et al. 2012] as introduced above for the well founded logical ontologies, or simply Bag-Of-Word (TF-IDF) overlapping measures in lexical based extensional models.

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 C_T ($x \in C_{X,T}^{\mathfrak{T}_2}$) 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) ($x \notin C_{X,T}^{\mathfrak{T}_2}$), or even it might be classified under another concept.

Definition 9: ***Cross-Lingual correspondence (Semantonym)***
Given two lexicalized concepts $c_S \epsilon \mathcal{O}s$, and $c_T \epsilon\, \mathcal{O}_T$ in $\mathcal{O}_S$ and $\mathcal{O}_T$ respectively, and lingualized 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.

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 *semantonyms* 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 s*emantonym* 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.

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

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 *classificational 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.

To be more formal, we present the *extension of a concept*, recalling that the set $C_{X,\mathcal{L}}^{\mathfrak{T}_1} = \{x \in X | x^{\mathfrak{T}_1} \in C_{\mathcal{L}}^{\mathfrak{T}_1}\}$ represents the subset of objects belonging to a shared context *X* classified under the lexicalized concept $C_{\mathcal{L}}$ 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:

1. *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*).
2. *corpus-based*: it is a corpus of documents, where each term in a given document can 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.

VII. CONCLUSION & FUTURE WORK

The cross-lingual mapping process goes beyond simply mapping concepts considering only the lexical dimension (considering only the labels), but we should take into consideration the use of concepts for classifying objects in a given context; the use normally varies; based on the community and the culture that ontology belongs to.

In this deliverable we have proposed a foundation for cross-lingual ontology matching framework by extending the mono-lingual mapping semantics to a cross-lingual one, taking into account ontology lexicalization. We used the classification based approach that has proved its usefulness in the ontology mapping domain. Particularly, we were inspired by the work presented in [Atencia et al. 2012] that proposed a formal semantics for weighted ontology mappings, based on a classification of interpretation for logically founded ontologies.

Several issues have to be resolved to confront the cross-lingual case; the mapping semantics of classification based interpretation needs to be extended to consider lexicalization in both logical as well as lexical ontologies; the impact

of different types of objects in the domain of interpretation needs to be studied; and an interesting subject is to study the mapping among mixed types (lexical and logical) of interpretations based on the variation of concept extensions. Finally the transitivity of the mappings, have to be framed within the classification-based semantics approach.

This deliverable has provided a first proposal of cross-lingual ontology matching framework was prepared. However, a complete definition of the semantonym, as well as validating the proposed mapping semantics by means of the alignment between Arabic Ontology and WordNet is out of the scope of this report and should be the target and focus for providing a final solution in the next deliverable (D2.3, M34) as well as in future research collaborations.

VIII. REFERENCES

[1] Manuel Atencia, Alexander Borgida, Jerome Euzenat, Chiara Ghidini, and Luciano Serafini. A formal semantics for weighted ontology mappings. In International Semantic Web Conference (1), pages 17-33, 2012.
[2] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors. Schema Matching and Mapping. Springer, 2011.
[3] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web, May 2001.
[4] Gosse Bouma. Cross-lingual dutch to english alignment using eurowordnet and dutch wikipedia. In OM, 2009.
[5] Paolo Bouquet, Marc Ehrig, Jrme Euzenat, Enrico Franconi, Pascal Hitzler, Marc Ehrig (u. Karlsruhe, Jrme Euzenat (inria, Enrico Franconi (fu, Pascal Hitzler (u. Karlsruhe, Markus Krtzsch (tu Dresden, Sergio Tessaris (fu Bolzano, Contact Person Dieter Fensel, and Contact Person Alain Leger. D2.2.1 specification of a common framework for characterizing alignment, technical report, knowledge web, 2004.
[6] 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.
[7] Michelle Cheatham. Mapsss results for oaei 2012, 2012.
[8] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. SIGMOD Rec., 35(3):34-41, September 2006.
[9] Philipp Cimiano, Elena Montiel-Ponsoda, Paul Buitelaar, Mauricio Espinoza, and Asuncion Gomez-Perez. A note on ontology localization. Applied Ontology, 5(2):127-137, 2010.
[10] William W. Cohen, Pradeep D. Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In IIWeb, pages 73-78, 2003.
[11] 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.
[12] Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In CIKM, pages 513-522, 2009.
[13] Gerard de Melo and Gerhard Weikum. Constructing and utilizing wordnets using statistical methods. Language Resources and Evaluation, 46(2):287-311, 2012.
[14] AnHai Doan and Alon Y. Halevy. Semantic integration research in the database community: A brief survey. AI Magazine, 26(1):83-94, 2005.
[15] Cassia Trojahn dos Santos, Paulo Quaresma, and Renata Vieira. An api for multi-lingual ontology matching.In LREC, 2010.
[16] Jerome Euzenat. Algebras of ontology alignment relations. In International Semantic Web Conference, pages 387-402, 2008.
[17] Jerome Euzenat and Pavel Shvaiko. Ontology matching. Springer, 2007.
[18] Richter F. Ding: Germany-English Dictionary, 1.5. 2007.
[19] Christiane Fellbaum. Wordnet: An electronic lexical database. Cambridge, MA. MIT Press,1998.
[20] Bo Fu, Rob Brennan, and Declan O'Sullivan. Cross-lingual ontology mapping - an investigation of the impact of machine translation. In ASWC, pages 1-15, 2009.
[21] 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.
[22] 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.
[23] Anika Gro, Michael Hartung, Toralf Kirsten, and Erhard Rahm. Gomma results for oaei 2012, 2012.
[24] Walid Hassen. Medley results for oaei 2012, 2012.
[25] Sven Hertling and Heiko Paulheim. Wikimatch - using wikipedia for ontology matching. In Seventh International Workshop on Ontology Matching (OM 2012), 2012.
[26] Jakob Huber, Timo Sztyler, Jan Noessner, and Christian Meilicke. Codi results for oaei 2011, 2011.

[27] Antoine Isaac, Lourens Van Der Meij, Stefan Schlobach, and Shenghui Wang. An empirical study of instance-based ontology matching. In Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, ISWC'07/ASWC'07, pages 253-266, Berlin, Heidelberg, 2007. Springer-Verlag.
[28] Ernesto Jimenez-Ruiz, Bernardo Cuenca Grau, and Ian Horrocks. Logmap results for oaei 2012, 2012.
[29] Jason J. Jung. Ontological framework based on contextual mediation for collaborative information retrieval. Inf. Retr., 10(2), April 2007.
[30] 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.
[31] Konstantinos Kotis, Artem Katasonov and Jarkko Leino, Automsv2 results for oaei 2012, 2012a.
[32] Konstantinos Kotis, Artem Katasonov, and Jarkko Leino. Ase results for oaei 2012, 2012b.
[33] Anita C. Liang and Margherita Sini. Mapping agrovoc and the chinese agricultural thesaurus: Definitions, tools, procedures. The New Review of Hypermedia and Multimedia, 12(1):51-62, 2006.
[34] Feiyu Lin and Andrew Krizhanovsky. Multilingual ontology matching based on wiktionary data accessible via sparql endpoint. In RCDL, pages 1-8, 2011.
[35] Feiyu Lin and Kurt Sandkuhl. A survey of exploiting wordnet in ontology matching. In IFIP AI, pages 341-350, 2008.
[36] Veronique Malaise, Antoine Isaac, Luit Gazendam, Telematica Instituut, and Hennie Brugman. Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), Prague, Czech Republic, page 57 64, 2007.
[37] George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In Pro. of the workshop on Human Language Technology, HLT '93, pages 303-308, Stroudsburg, PA, USA,1993. Association for Computational Linguistics.
[38] Malgorzata Mochol. The methodology for finding suitable ontology matching approaches. PhD thesis, Freie Universitat Berlin, Germany, 2009.
[39] Miklos Nagy, Maria Vargas-Vera, and Piotr Stolarski. Dssim results for oaei 2009. In OM, 2009.
[40] 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.
[41] Grace Ngai, Marine Carpuat, and Pascale Fung. Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In COLING, 2002.
[42] DuyHoa Ngo and Zohra Bellahsene. Yam++ : A multi-strategy based approach for ontology matching task. In EKAW, pages 421-425, 2012.
[43] Heiko Paulheim. Wesee results for oaei 2012, 2012.
[44] Pavel Shvaiko. Iterative schema-based semantic matching. PhD thesis, University of Trento, Italy, 2006.
[45] Pavel Shvaiko and Jerome Euzenat. Ontology matching: State of the art and future challenges. IEEE Trans. Knowl. Data Eng., 25(1):158-176, 2013.
[46] Lichao Song. The role of context in discourse analysis. Journal of Language Teaching and Research, 1:876-879,2010.
[47] 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.
[48] Raquel Trillo, Jorge Gracia, Mauricio Espinoza, and Eduardo Mena. Discovering the semantics of user keywords. J. UCS, 13(12):1908-1935, 2007.
[49] Teun A. VAN DIJK. Discourse context and cognition. discourse studies. Discourse Studies, 8:159-177, 2006.
[50] Shenghui Wang, Antoine Isaac, Balthasar A. C. Schopman, Stefan Schlobach, and Lourens van der Meij. Matching multi-lingual subject vocabularies. In ECDL, pages 125-137, 2009.
[51] Zhichun Wang, Xiao Zhang, Lei Hou, Yue Zhao, Juanzi Li, Yu Qi, and Jie Tang. Rimom results for oaei 2010,2010.
[52] Zhang X., Zhong Q., Li J.and Tang J., Xie G., and Li H. Rimom results for oaei 2009, 2009.

C. Conclusions

This deliverable has given an overview of the most important research set up activities undertaken by the SIERA consortium in the first 18 months of the period.

So far the main achievements can be summarized by the following points:

- The missing Arabic tools were specified, they were designed and finalized to work with minimal installation on MICHAEL search system, that is they are ready for demo testing. The activity results currently depends on the results of integration which will be provided by MICHAEL developers after testing the tools. Please note that the tools that need further research and development will be highlighted after the conduction of the integration test, those tools can be jointly tackled by the project partners in the future.
- MICHAEL's thesaurus in SKOS format was extended with Arabic.
- SKOS format of Bethlehem Thesaurus was produced using the TMP environment after defining a domain ontology for Bethlehem data.
- The Arabic cultural objects and named entities are being processed by OKKAM to be connected then to MICHAEL.
- The Arabic Ontology Top Levels were mapped to KYOTO, the achieved mappings will be reassessed and extended, where each concept in KYOTO will be mapped to its Arabic equivalent.
- A framework for mapping between WordNet and Arabic Ontology were defined. The mapping and the evaluation of the proposed framework will be the target of the future work.