Towards a Public Multilingual Knowledge Management Infrastructure for the European Digital Single Market

Peter Schmitz, Enrico Francesconi*, Najeh Hajlaoui, Brahim Batouche

Publications Office of the European Union, Luxembourg
* Publications Office of the European Union, Luxembourg and ITTIG-CNR, Florence, Italy

{Peter.Schmitz, Enrico.Francesconi}@publications.europa.eu
{Najeh.Hajlaoui, Brahim.Batouche}@ext.publications.europa.eu

Abstract.
This paper describes the first phase of the Public Multilingual Knowledge Management Infrastructure (PMKI) ISA2 project. PMKI is meant to support enterprises, in particular the language technology industry, as well as public administrations, with multilingual tools able to improve cross border accessibility of digital services. In particular it aims to create a set of tools and facilities, based on Semantic Web technologies, for establishing semantic interoperability between multilingual lexicons. A comparative study among the main data models for representing lexicons and recommendations for the PMKI service are reported. Moreover, the expected synergies with other programs of the EU institutions, as far as systems interoperability and machine translation solutions, are discussed.

Keywords: Multilingual Language Resource Infrastructure, standard representation, data model, interoperability.

1 Introduction

Linguistic and cultural diversity is one of the cornerstones that the European Union is built upon. It represents a treasure that needs to be preserved but, at the same time, a challenge to face considering the barriers which come along and which have to be removed in order to guarantee a space of freedom, justice and democracy for millions of European citizens. The EU economy is one of the main areas that can benefit from overcoming such barriers and can exploit the globalization of the markets, which is sped up by a rapid digital transformation of the society. In this respect the creation of a Digital Single Market[1] for the EU is one of the main priorities of the European

[1] https://ec.europa.eu/commission/priorities/digital-single-market_en

E. Francesconi is author of Section 1, B. Batouche of Section 2 and 3, N. Hajloui of Section 4, P. Schmitz of Section 5

CEUR-WS.org/Vol-1899/OntoLex_2017_paper_1.pdf

Commission aimed to provide better online access to digital goods and services, guarantee an environment where digital networks and services can prosper, as well as make "digital" as a driver for growth.

The pilot project Public Multilingual Knowledge Infrastructure (PMKI), launched within the ISA2 program, aims to represent a contribution to overcome the barriers hampering an effective cross-border exploitation of Digital Single Market services, in particular language and semantic barriers.

The project aims to create a set of tools and facilities, based on semantic Web technologies, aimed to support enterprises, in particular the language technology industry, as well as public administrations with multilingual tools in order to improve cross border accessibility of e-commerce solutions and public services, helping to build the Connecting Europe Facility Automated Translation (CEF.AT) Platform[2] - a common building block implemented through the CEF programme.

One of the main objectives of PMKI is therefore to establish semantic interoperability between digital services, which, in practical terms, means overcoming language and technical barriers on the Web by creating multilingual vocabularies and lexicons, as well as establishing links between them. This will improve the ability of digital systems to exchange data with unambiguous, shared meaning, thus supporting the accessibility of services and goods offered through the Internet.

This paper describes the first phase of the PMKI project, which consists in a feasibility study about the implementation of mapping facilities and relations between multilingual lexicons based on semantic Web technologies. It is organized as follows: in Section 2 an overview of the main objectives of the project is given; in Section 3 comparative studies is illustrated about Semantic Web standards for representing multilingual lexicons and describing their relations, as well as possible platforms for managing lexicons; in Section 4 the PMKI relationships with the CEF (Connecting Europe Facility) program of the European Commission are discussed. Finally in Section 5 some conclusions are reported.

2 PMKI project

2.1 Presentation

The objective of PMKI is to implement a proof-of-concept infrastructure to expose and to harmonize internal (European Union institutional) and external multilingual lexicons aligning them in order to facilitate interoperability. Additionally the project aims to create a governance structure to extend systematically the infrastructure by the integration of supplementary public multilingual taxonomies/terminologies.

PMKI is a pilot project to check the feasibility and to prepare a road map to convert such proof of concept into a public service.

[2] https://ec.europa.eu/digital-single-market/en/automated-translation

2.2 Why such platform

The need to have a public and multilingual platform that can play the role of a hub to collect and to share language resources in standardised formats is essential to guarantee semantic interoperability of digital services. For instance, such platform is missing in CEF.AT, while it would provide an advantage for the development of machine translation systems. In particular it can provide alignments of domain specific terminologies for developing specific-domain translation systems (tender terminology, medical terminology, etc.).

A platform like PMKI may represent a one-stop-shop harmonized multilingual lexicons repository at European level.

Contrary to the European Language Resource Coordination (ELRC[3]) action, which aims to identify and gather language and translation data, PMKI platform aims first to harmonize multilingual language resources making them interoperable, then to integrate supplementary public multilingual taxonomies/terminologies in a standardized representation. That is why we need first to define a 1) sophisticated standard representation that will be used with respect to a 2) defined core data model (in case with extensions) under 3) an adequate architecture.

These three requirements are respectively analyzed and detailed in three first analysis phases of the project:

- Analysis of existing relevant standards for the representation of lexicons that will be made available on the PMKI platform;
- PMKI core data model and extensions (based on the standard representation that is recommended as result of the previous analysis);
- Analysis of available platforms for managing lexicons.

Interoperability is one of the main features of the PMKI platform; such platform will provide support to develop multilingual tools such as machine translation, localization, search etc. For instance for the machine translation tool, interoperable translation data is a factor of success to improve the quality mainly for under-resourced languages.

[3] ELRC2: the European Language Resource Coordination action launched by the European Commission as part of the CEF.AT platform activities, to identify and gather language data across all 30 European countries participating in the CEF programme. This will be followed up by actions concerning the setting up of a repository of language resources for CEF AT and further data collection and awareness actions in the context of calls for tenders and calls for proposals for which the selection procedure is still ongoing. More information can be found here: http://www.lr-coordination.eu/

3 Analysis and comparative studies

3.1 Requirements collection

To select the most appropriate standard for the PMKI project, we have collected our requirements and refer them to the available standards for lexical resources and their enrichment. In particular we analyzed SKOS (Alistair & all, 2005), WordNet, Lemon (Villegas, 2015) and OntoLex (Bosque-Gil and all, 2015). The collected requirements are associated to the type of lexical resources which PMKI aims to deal with and to the type of their possible lexical enrichment.

3.1.1. Type of lexical resources

In our analysis we have identified the PMKI requirements covered by the following resource types: (1) controller vocabularies, (2) glossaries, (3) lexicons, (4) thesauri, (5) taxonomies and (6) semantic networks. The types (1-3) concern mainly the linguistic community; the types (4-5) concern mainly the librarian community and the type (6) concerns both communities.

For each type of resources, we define our requirements as a projection of the resources type. These resources are listed in Tab. 1, where the second colon describes the projected requirements.

Resource type	Requirements
Controlled vocabularies: list of terms or sentences that contribute to the homography and support disambiguation	Description of multiword expression
Glossary: list of words with definitions	Definition of terms
Thesaurus: controlled vocabulary with definitions, properties and relations between terms/concepts	Description of relation between terms independently of domain/context
Lexicon (lexical-Semantic databases): set of terms and linguistic relations between them	Definition of relation between terms/concepts, depending on domain/context
Taxonomy: hierarchical classification of concepts	Hierarchical classification of concepts which depends on the domain
Semantic Network: knowledge representation of nodes (objects or classes) and relations, usually organized in a direct graph. It does not support formal constraints and can even not be a shared knowledge as the ontologies are.	Conceptualization of the term sense by linking it with semantic resources (domain ontology)

Table 1 List of requirements and related lexical resource types

Table 2 gives a comparison between the main linguistic data models available in literature according to the type of resources which cover the PMKI requirements.
From such a comparison it is evident that while SKOS is mainly devoted to describing thesauri and taxonomies, Lemon is mainly oriented to describe vocabularies, glossaries, lexicons and semantic networks of term, while Ontolex ontology is able to describe all the selected types of lexicons.

	Vocabulary	**Glossary**	**Lexicon**	**Thesaurus**	**Taxonomy**	**Semantic Network**
SKOS	No	No	No	Yes	Yes	No
WordNet	No	Yes	No	No	No	No
LEMON	Yes	Yes	Yes	No	No	Yes
OntoLex	Yes	Yes	Yes	Yes	Yes	Yes

Table 2 Comparison between data models and covered lexical resources

3.1.2. Lexical enrichment

From the lexical enrichment point of view, the main requirements for the PMKI lexical resources are: (a) hierarchical classification of concepts, (b) linguistic variation of terms (c) sense of terms, (d) multiword expression as phrases, (e) semantics of terms.

Table 3 illustrates how the analysed standards are able to deal with such requirements. Also from this comparison, OntoLex model is the one best fitting the requirements of the PMKI project.

	Hierarchical concepts	**Linguistic variation**	**Word Sense**	**Phrase**	**Semantics of terms**
SKOS	Yes	No	No	No	No
WordNet	Yes	No	Yes	No	Yes
LEMON	No	Yes	Yes	Yes	Yes
OntoLex	Yes	Yes	Yes	Yes	Yes

Table 3 Comparison between data models and lexical enrichment requirements

3.2 Data model

From the comparative study of the available data models, it emerged that Lemon-OntoLex[4] standard covers the PMKI requirements, therefore it represents the most promising candidate to become the data model for PMKI.

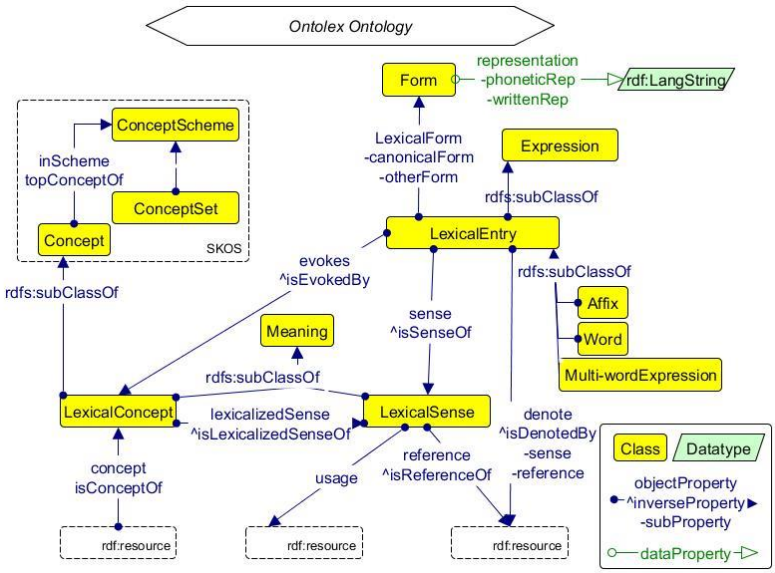

Figure 1 Ontolex core ontology

Ontolex is based on around the idea of separation between the lexical and the ontological layer. Additionally it is modelled to be modular and extensible in giving a hub for connecting lexical and terminological description elements with ontologies. In particular the sub-ontology (modules) of OntoLex are: Ontolex core[5] (illustrated in Figure 1), syntax/semantic[6], decomposition[7], variation/transformation[8], metadata (Lime)[9].

OntoLex describes the meaning of a word by reference to the data model, a lexical entry may be associated with a lexical concept which is sub class of skos:concept. Lexical concept represents the semantic pole of linguistic units and is the mentally instantiated abstraction which language users derive from conceptions. We consider

[4] https://www.w3.org/2016/05/ontolex/
[5] http://www.w3.org/2004/02/skos/core#
[6] http://www.w3.org/ns/lemon/synsem#
[7] http://www.w3.org/ns/lemon/decomp#
[8] http://www.w3.org/ns/lemon/vartrans#
[9] http://www.w3.org/ns/lemon/lime#

the abstraction of lexical concept as skos:concept an interesting feature because it allows to consider a topic and its hierarchy, which is a feature not covered by LEMON.

3.3 Platform for managing language resources

As said, the objective of PMKI project is to promote a Digital Single Market in the EU. To this aim the alignment of multilingual lexicons dataset are essential for cross-border accessibility of public administration services and e-commerce solutions. In order to offer PMKI services, a knowledge management platform including functionalities and features, such as editing of lexicons, interoperability between linguistic models, alignments of lexical resources, import/export of dataset, and merging of dataset, is necessary. Based on Semantic Web technologies, PMKI will exploit and improve the existing datasets with semantic hyperlinking.

In literature one of the most advanced platforms able to manage lexical resources is the editor VocBench (Stellato & al., 2011), similarly a Web application for establishing and managing language resources and their interoperability is BabelNet.

4 Possible synergies with CEF.AT

The Connecting Europe Facility (CEF)[10] was established by Regulation (EU) N°1316/2013 of the European Parliament and of the Council of 11 December 2013 (CEF Regulation)[11]. It determines the conditions, methods and procedures to provide the European Union (EU) financial assistance to trans-European networks in order to support trans-European networks and infrastructures in the sectors of transport, telecommunications and energy.

CEF.AT is the Automated Translation platform to be developed as part of the Connecting Europe Facility (CEF) Programme; it is actually one of its building blocks. The starting point of the CEF.AT is the Machine Translation at the European Commission (MT@EC) system, which is a translation service available since 2013 to EU Institutions, Member State administrations and a number of EC Information systems and online services.

CEF.AT is a secure and adaptable automated translation platform that plugs into any Digital Service Infrastructure[12] (DSI) such as eHealth, eProcurement, eJustice, etc. and makes it multilingual. CEF.AT will eventually offer tools and services for the multilingual enablement of public services. It is a key enabler for cross-border public digital services and public administration within the Digital Single Market.

---

[10] http://ec.europa.eu/digital-agenda/en/connecting-europe-facility

[11] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R1316&from=EN

[12] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Sector+Specific+DSI

4.1 PMKI and MT

PMKI output can be integrated as a module or service to the Connecting Europe Facility Automated Translation platform (CEF.AT).

CEF.AT platform could benefit from the PMKI project to support the implementation of the necessary multilingual tools and features (machine translation, localization and multilingual search) offered to digital services (including the sectoral DSIs) and public institutions at the EU and member states level, contributing to disambiguation, domain identification, facilitation of cross-lingual search and machine translation.

The PMKI strategy consists to find a sophisticated core data model based on a standard representation for multilingual taxonomies and terminologies and its implementation for the representation of the data in a public platform. Such standard representation will make it possible for different stakeholders to manage the development and evolution of their own data on an individual base, but at the same time enable interoperability through alignment and linking can be useful both as a data/content resource and as a tool/service.

PMKI service can be used for better leveraging both the new data that are collected to support customization and adaptation in CEF.AT, and the data currently used by MT@EC. It can also be used to attract data providers who wish to leverage their own data and users/DSIs who may wish to use it for their own needs (for example visualization, search, domain identification and classification, etc.)

PMKI as a digital service will play a vital role in the collection and the share of data across borders and domains. It will help European institutions and public administrations exchange resource data across language barriers in the European Union. For instance, it will help the CEF.AT to make all Digital Service Infrastructures (DSI) multilingual by playing the role of a hub that offers a common infrastructure to enable multilingual services. In order to facilitate interoperable services as foreseen by the Digital Agenda for Europe, PMKI aims to facilitate first interoperability between language resources proposing a Multilingual Knowledge Infrastructure which:

- will represent a new data hub to collect and share data resources in standardized formats;
- will be reused for CEF.AT to provide bilingual data to build/reinforce machine translation systems, in particular on specific domain systems (tender terminology, medical terminology, etc.);
- will be the main point of interoperable language resources coordination.

The actual CEF.AT platform produces only a very limited number of direct translations. Direct translations are generally from English/French/German (the 3 main working languages in the EU institutions) to all EU languages and from all EU languages to English/French/German. The remaining combinations are indirect translations via one of the three main languages. PMKI language resources such as EuroVoc, which is already multilingual (24 languages), will allow the production of new parallel corpora for such languages to train new direct MT systems.

4.2 **Support for specific domain data**

A sublanguage is a subset of the language (Harris, 1970) identified with a particular semantic domain (or a family of domain) (Kittredge, 1978), (Kittredge, 1982).

(Hajlaoui & Boitet, 2008) showed that, in case of very small sublanguages, SMT may be sufficient quality, starting from a corpus 100 to 500 smaller than for the general language. The corpus must obviously reach a critical size to allow reliable statistical treatment. SMT approach works very well for restricted domains with little or no human revision, for example the rules-based TAUM-METEO system is purposely developed for the weather service in Canada to provide weather forecasts in French and English. Quality offered by machine translation could be very high under certain preconditions related to the type of domain sublanguage:

- Lexical convergence of the domain-sublanguage: the vocabulary used to prepare the SMT system should cover the domain.
- Grammatical convergence: the training data should cover all the grammatical variations that can be used with the domain-sublanguage.

In this and similar domains, the quality of translations theoretically can and practically must be gradable. The performance of a specific SMT system is proportional to the coverage of the domain. The coverage is usually reached after a certain size of training data. A relevant parallel corpora related to a given domain is usually difficult to obtain. The idea is therefore to use PMKI as a filter to produce such specific domain data. For instance users, such as specific-domain MT developers, can query a document collection in a specific language for a specific domain. Using CEF.AT users can translate his specific query into 23 languages. Then, by selecting a list of concepts in the languages of his interest he will obtain the list of filtered documents in multiple languages that concern the initial specified domain thanks to the multi-lingual and multi-collection concepts associated to the documents.

4.3 **Support for Machine Translation (MT) and Translation Memory (TM)**

Actually language resources are the "raw" material for statistical machine translation and Neural Machine Translation approaches. The EU spends lots of efforts and money to develop and maintain lexicons for developing services to be used in the CEF programme. PMKI platform can be an effective contribution to the MT activities, providing a network of interoperable multilingual lexicons. The cross-lingual and cross-collection retrieval services offered by PMKI will be able to:

- Improve the quality of MT/TM by adding new translation data.
- Improve MT/TM quality for under-resourced languages pairs.
- Facilitate the development of new direct MT systems for languages pairs where indirect translations are actually used.

PMKI can represent a “one-stop shop” for accessing lexicons, as well as providing tools for harmonizing their technological formats, localizing them and making them interoperable with other lexicons.

PMKI can provide parallel data that can be processed for MT and TM as well in order to maximize the probability of reuse of previous human translations.

5 Conclusion

In this paper the preliminary results obtained within the PMKI project for implementing interoperability solution of lexical resources is presented. In particular the main Semantic Web standards available in literature for representing lexicons have been identified and their characteristics analyzed. For the ability of describing lexical components in different languages using LEMON, the related concepts and their mapping relations using SKOS, the Ontolex standard resulted as the preferred model to be adopted as reference for the PMKI platform.
The possible benefits of PMKI in the field of Machine Translation and Translation Memory have been discussed.

The next phase of the project will provide an evaluation of different mapping algorithms and propose a technical infrastructure for the implementation and maintenance of lexical resources and their interoperability.

6 References

- Alistair Miles, Brian Matthews, Michael Wilson, Dan Brickley, (2005) SKOS Core: Simple knowledge organisation for the Web, DCMI
- Stellato & al., (2011) VocBench: a Web application for Collaborative development of multilingual Thesauri, adfa, p. 1.
- Bosque-Gil & al., (2015) Applying the OntoLex Model to a Multilingual Terminological Resource, ESWC 2015
- Hajlaoui, N., & Boitet, C. (2008). TA statistique à petits corpus pour de petits sous-langages. Proc. ToTh 2008 : Conférence sur la Terminologie & Ontologie : Théories et Applications, (p. 15). France - Annecy.
- Harris, Z. (1970). Mathematical structures of language. Mathematical Gazette. Vol. 54(388), 173-174.
- Kittredge, R. (1978). Textual cohesion within sublanguages: implications for automatic analysis and synthesis. Proc. Coling-78. Vol. 1/1. August 14-18. Bergen, Norvège.
- Kittredge, R. (1982). Variation and Homogeneity of Sublanguages. In Sublanguage -Studies of Language in Restricted Semantic Domains., (p. 20). Walter de Gruyter. Berlin / New York.
- Koehn Philipp et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In 45th Annual Meeting of the ACL, Demonstration Session, pages 177–180, Prague
- Villegas, M. & Bel, N. (2015). PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web*, 6, 363-369.