Dataset Alignment and Lexicalization to Support Multilingual Analysis of Legal Documents

Armando Stellato*, Manuel Fiorelli*, Andrea Turbati*, Tiziano Lorenzetti*
Peter Schmitz+, Enrico Francesconi+§, Najeh Hajlaoui+, Brahim Batouche+

* ART Group, Dept. of Enterprise Engineering
University of Rome Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
stellato@uniroma2.it
{fiorelli, turbati}@info.uniroma2.it
tiziano.lorenzetti@gmail.com

+ Publications Office of the European Union, Luxembourg
Publications Office of the European Union, Luxembourg
{Peter.SCHMITZ, Enrico.Francesconi}@publications.europa.eu
{Najeh.HAJLAOUI, Brahim.BATOUCHE}@ext.publications.europa.eu

§ Institute of Legal Information Theory and Techniques (ITTIG)
Consiglio Nazionale delle Ricerche (CNR)
Via dei Barucci 20 - 50127 Florence, Italy

Abstract. The result of the EU is a complex, multilingual, multicultural and yet united environment, requiring solid integration policies and actions targeted at simplifying cross-language and cross-cultural knowledge access. The legal domain is a typical case in which both the linguistic and the conceptual aspects mutually interweave into a knowledge barrier that is hard to break. In the context of the ISA^2 funded project "Public Multilingual Knowledge Infrastructure" (PMKI) we are addressing Semantic Interoperability at both the conceptual and lexical level, by developing a set of coordinated set of instruments for advanced lexicalization of RDF resources (be them ontologies, thesauri and datasets in general) and for alignment of their content. In this paper, we describe the objectives of the project and the concrete actions, specifically in the legal domain, that will create a platform for multilingual cross-jurisdiction accessibility to legal content in the EU.

1 Introduction

The construction of the European Union is one of the most political success stories of the last decades, able to guarantee a space of freedom, justice and democracy for millions of European citizens, based on the free exchange of people, information, goods and services.

However, complex, multilingual and multicultural as Europe is, it cannot rely on political success and good intentions alone: the objectives of its unification must be underpinned by solid integration policies and targeted actions, considering and dealing with the heterogeneities that lay at the basis of the foundation of EU itself.

The legal domain is an emblematic example of this heterogeneity: while united under common goals and ethics, each of the Member States retains its own laws and regulations. These need to be aligned to the common directions and indications provided by the EU Parliament, while keeping their independence and bindings to the constitutions characterizing each nation. The differences are not technically limited to the regulations per se, being the whole fabric of knowledge bond to the cultural and societal heritage of a nation. For instance, the French concept "tribunaux administratifs" cannot be translated in English as "administrative tribunals". The English word for "tribunaux" in fact is "courts" while the "administrative tribunals" are administrative commissions which are comparable, mutatis mutandis, to the French "autorités administratives indépendantes" [1]. There is however, as this example shows, a linguistic problem as well, as it is important that the reached semantic consensus on recognized similarities and affinities be available and accessible in different languages.

In such a scenario, the European digital eco-system should be made ready to support seamless and cross-lingual access to Member States' legislations, accounting for their differentia as well as their relatedness under the common umbrella of the EU.

With this objective to pursue, and in a broader context including, but not limited to, the domain of jurisprudence and law, in 2010 the EU defined the so-called European Interoperability Framework, namely a set of recommendations and guidelines to support the pan-European delivery of electronic government services. This framework aims at facilitating public administrations, enterprises and citizens to interact across borders, in a pan-European context. Such guidelines cover different aspects of social, commercial and administrative relations among different European actors, like multilingualism, accessibility, security, data protection, administrative simplification, transparency, reusability of the solutions.

One of the main objectives of such guidelines is to establish semantic interoperability between digital services, having the potential to overcome the barriers hampering their effective cross-border exploitation, which means making information exchange not only understandable by humans but also understandable and processable by machines, as well as establishing correspondences between concepts in different domains and languages, or represented in different digital tools (like controlled vocabularies, classification schemas, thesauri).

In the context of the Public Multilingual Knowledge Infrastructure (PMKI), a project funded by the ISA[2] programme[1] with the aim to overcome language barriers within the EU by means of multilingual tools and services, we are addressing Semantic Interoperability at both the conceptual and lexical level, by developing a set of coordinated set of instruments for advanced lexicalization of RDF resources (be them ontologies, thesauri and datasets in general) and for alignment of their content.

[1] https://ec.europa.eu/isa2/

In this paper, we will show how the realization of such an objective will enable seamless, multilingual, cross-legislative retrieval and analysis of legal content, and will show how the PMKI project will contribute to such a vision by detailing it objectives and milestones. The rest of the paper is organized as follows: section 2 provides more motivations for our effort and describes use case scenarios. In section 3 a brief overview on the evolution of models for representing lexical resources is given. Section 4 introduces the PMKI project while section 5 details the actions of the project and their outcomes in the legal domain. Section 5 concludes the paper.

2 Use-Case Scenarios

There are several scenarios in the management and access to legal content that would benefit from a thorough approach to conceptual and lexical integration.

2.1 Semantic Integration

As shown in [2], a typical use case for the adoption of legal ontologies is their application to specialist domains (e.g. industry standards), which in turn opens up different interaction possibilities between the legal knowledge (e.g. norms and regulations) and the one pertaining to the specialized domain (e.g. a domain ontology related to the aforementioned industry standards). Alignments between legal ontologies and specialized domain ontologies should be considered as precious resources per se, as well as the systems and frameworks that support the creation of such artifacts.

Semantic integration between analogous legal knowledge resources developed in different countries is also important, in order to facilitate the understanding of alien concepts through those closer to one's own culture or, at least, through general shared conceptualizations. For instance, the Italian thesaurus TESEO[2] (TEsauro SEnato per l'Organizzazione dei documenti parlamentari: Senate Thesaurus for the organization of parliamentary documents) is a classification system, originally developed by the Italian Senate and now used in the most relevant databases of the Senate, Chamber of Deputies and of some regions of Italy. Even in the multilingual environment characterizing the EU, the monolingual TESEO (its concepts are expressed in Italian only) keeps its relevance due to its tight connection to the Italian law regulation system and culture (TESEO includes a mix of specific legal concepts and more general topics). It is thus an irreplaceable resource for semantically indexing information from the above-mentioned Italian data and document bases. At the same time, aligning TESEO to other resources, such as the EU's multilingual thesaurus EuroVoc[3], the multilingual thesaurus of the EU, allows for cross-cultural and cross-lingual (EuroVoc is available in 26 languages, chosen from those spoken by EU Member States and candidate countries) mediated access to any content indexed through it. While EuroVoc obviously lacks the specificities that TESEO can offer to the Italian interested user, it still provides a best-mediated access modality, universally accepted and officially adopted within the EU.

[2] https://www.senato.it/3235?testo_generico=745
[3] http://eurovoc.europa.eu/

2.2 Natural Language Understanding

Another very important scenario (which, in turn, includes a plethora of use cases) concerns the identification and extraction of relevant information. The identified information can be exploited in a variety of tasks, such as:

Annotation of document corpora. In the legal domain law references, articulated in proper structures (e.g. law, article, paragraph), are important indexing elements for documents. Being able to search corpora by including explicit constraints on mentioned regulations is a powerful feature for legal search engines. For instance, a lawyer could access to the full list of trials registered for a given court, and extract all judgements that include a mention of a given law (or a set of laws), eventually specifying portions of them. However, discovering references is not trivial, as it implies both being able to parse the structure of a law mention (a mix of natural language processing capabilities and of background knowledge about the existing laws is required) and being able to recognize the so called "popular expressions" for referring to these laws. For instance, in Italy the expression "Bossi-Fini" is a specific term referring to the law n° 189 of 30th July 2002, establishing policies about immigration and employment of migrants. This popular expression is originated by the names of the first signatories, Gianfranco Fini and Umberto Bossi, at that time vice-president of the cabinet and minister for institutional reforms and devolution respectively, and is often adopted even in specialized literature. A proper lexicalization even – and actually, most importantly – in the same language of the country adopting that law is thus important in order to recognize the variety of expressions that refer to the same precise legal entity.

Knowledge Building. While the previous scenario deals with the identification of mentions of entities already defined and structured in specific areas of knowledge, it is also important to be able to build new knowledge by analyzing language content. The development of specialized domain ontologies can be partially automated by applying terminology extraction techniques to document corpora [3] in order to identify the entities that will be later elaborated into ontology classes and properties. The analysis of relations [4] in the text can help both the development of new knowledge as well as – when legal content is available – the production of alignments between legal and domain ontologies.

Cross-lingual recognition. The availability of terms in multiple languages allows for efficient retrieval of the same conceptual information in various languages. However, the analysis of language content (for any of the two tasks above) requires more fine-grained lexical background knowledge than just mere terminology. Being able to describe, in different languages, the single components forming compound terms, the several forms in which these can be declined/conjugated, their lexical variations etc.. is a necessary step which has to be carried on even for those languages different from the one spoken in the country where that knowledge originated.

It appears evident how all the tasks above would benefit from proper lexicalization of the knowledge resources involved, performed by adopting well established standards for the representation of lexical information and of the lexical-semantic interface with

ontologies. Re-use of existing resources modeled according to these standards should be also encouraged, to minimize the effort for lexicalizing knowledge resources. In the next section, we will provide an excursus over models for lexical resources that have been proposed in the last 20 years of research on (computational) linguistics.

3 State of the Art on Linguistic Resources and Language Representation

"The term linguistic resources refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems" [5].

Multiple efforts have been spent in the past towards the achievement of consensus among different theoretical perspectives and systems design approaches. The Text Encoding Initiative (www.tei-c.org) and the LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards) project [6] are just a few, bearing the objective of making possible the reuse of existing (partial) linguistic resources, promoting the development of new linguistic resources for those languages and domains where they are still not available, and creating a cooperative infrastructure to collect, maintain, and disseminate linguistic resources on behalf of the research and development community.

A popular resource which got a broad diffusion characterized by exploitation in both applications and scientific studies is WordNet [7,8]. Being a structured lexical database, presents a neat distinction between words, senses and glosses, and is characterized by diverse semantic relations like hypernymy/hyponymy, antonymy etc… Though not being originally realized for computational uses, and being built upon a model for the mental lexicon, WordNet has become a valuable resource in the human language technology and artificial intelligence. Due to its vast coverage of English words, WordNet provides general lexico-semantic information on which open-domain text processing is based. Furthermore, the development of WordNets in several other languages [9,10,11] extends this capability to trans-lingual applications, enabling text mining across languages.

A more recent effort towards achieving a thorough model for the representation of lexical resources is given by the Lexical Markup Framework [12]. LMF, which has obtained ISO standardization (LMF; ISO 24613:2008), can represent monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons, for both simple and complex lexicons, for both written and spoken lexical representations. The descriptions range from morphology, syntax, computational semantics to computer-assisted translation. The covered languages are not restricted to European languages but cover all natural languages. The range of targeted NLP applications is not restricted. LMF is able to represent most lexicons, including the above mentioned WordNet.

With the advent of the Semantic Web and Linked Open Data, a number of models have been proposed to enrich ontologies with information about how vocabulary elements have to be expressed in different natural languages. These include the Linguistic Watermark framework [10,11], LexOnto [12], LingInfo [13], LIR [14],

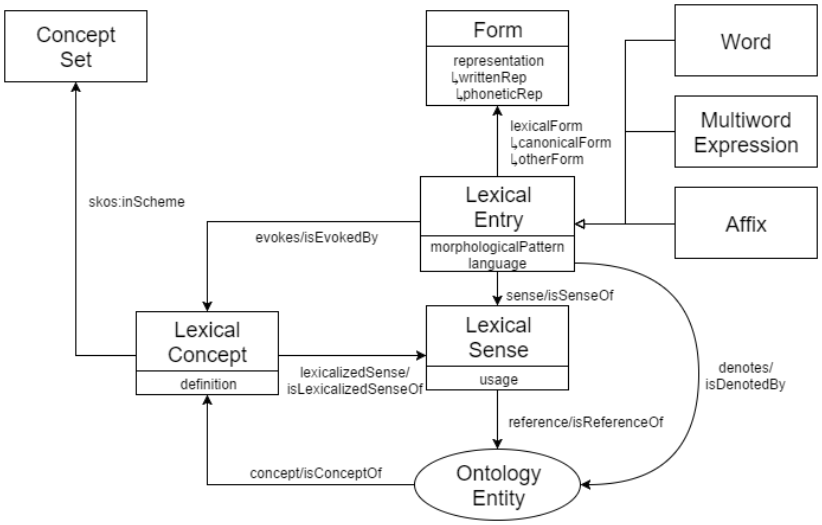

Fig. 1. The OntoLex-Lemon Model

LexInfo [1] and, more recently, *lemon* [15]. The *lemon* model envisions an open ecosystem in which ontologies[4] and lexica for them co-exist, both of which are published as data on the Web. It is in line with a many-to-many relationship between: i) ontologies and ontological vocabularies, ii) lexicalization datasets and iii) lexical resources. Lexicalizations in our sense are reifications of the relation between an ontology reference and the lexical entries by which these can be expressed within natural language. *lemon* foresees an ecosystem in which many independently published lexicalizations and lexica for a given ontology co-exist.

In 2012, an important community effort has been made to provide a common model for Ontology-Lexicon interfaces: the OntoLex W3C Community Group[5] was started with the goal of providing an agreed-upon standard by building on the aforementioned models, the designers of which are all involved in the community group.

The OntoLex-*lemon* [13] model (see Fig.1) developed by the OntoLex Community Group is based on the original *lemon* model, which by now has been adopted by a number of lexica [14,15,16,17], and as such was taken by the group as the basis for developing an agreed-upon and widely accepted model. The *lemon* model is based onto the idea of a separation between the lexical and the ontological layer following Buitelaar [18] and Cimiano et al [19], where the ontology describes the semantics of the domain and the lexicon describes the morphology, syntax and pragmatics of the words used to express the domain in a language. The model thus organizes the lexicon

[4] It would be more appropriate to adopt the term "reference dataset" (including thus also SKOS thesauri and datasets in general), to express data containing the logical symbols for describing a certain domain. In line with the traditional name OntoLex (and thus the ontology-lexicon dualism), we will however often refer to them with the term ontology

[5] `http://www.w3.org/community/ontolex/`

primarily by means of *lexical entries*, which are a word, affix or multiword expression with a single syntactic class (part-of-speech) to which a number of *forms* are attached, such as for example the plural, and each form has a number of *representations* (*string forms*), e.g. written or phonetic representation. Entries in a lexicon can be said to *denote* an entity in an ontology, however normally the link between the lexical entry and the ontology entity is realized by a *lexical sense* object where pragmatic information such as domain or register of the connection may be recorded.

In addition to describing the meaning of a word by reference to the ontology, a lexical entry may be associated with a *lexical concept*. Lexical concepts represent the semantic pole of linguistic units, mentally instantiated abstractions which language users derive from conceptions [20]. Lexical concepts are intended primarily to represent such abstractions when present in existing lexical resources, e.g. synsets for wordnets.

Finally, linguists have acknowledged [21] the benefits that the adoption of the Semantic Web technologies could bring to the publication and integration of language resources, thus denoting a convergence of interests and results rarely occurring before. A concrete outcome of this convergence is given by the Open Linguistics Working Group[6] of the Open Knowledge Foundation, which is contributing to the development of a LOD (Linked Open Data) (sub)cloud of linguistic resources, known as LLOD[7] (Linguistic Linked Open Data).

4 The PMKI Project

Public Multilingual Knowledge Infrastructure (PMKI) is launched as an ISA2 action to answer claims from the European Language Technology Community such as the multilingual extension of the Digital Single Market, the increase of the EU cross-border online service. It aims to provide support for the EU economy in particular to SMEs to overcome language barriers and to help to unlock the e-Commerce potential within the EU implementing the necessary multilingual tools and features and helping to build the Connecting Europe Facility Automated Translation (CEF.AT) Platform - a common building block implemented through the CEF programme.

The project aims to create a set of tools and facilities, based on Semantic Web technologies, aimed to support the language technology industry as well as public administrations, with multilingual tools in order to improve cross border accessibility of public administration services and e-commerce solutions. In practical terms, overcoming language barriers on the Web means creating multilingual vocabularies and language resources, establishing links between them as well as using them to support accessibility to services and goods offered through the Internet.

The objective of PMKI is to implement a proof-of-concept infrastructure to expose and to harmonise internal (European Union institutional) and external multilingual resources aligning them in order to facilitate interoperability. It could support the knowledge layer of the multilingual infrastructure for Europe. Additionally the project

[6] http://linguistics.okfn.org/
[7] http://linguistic-lod.org/llod-cloud

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 of the proposed solutions and to prepare the roadmap to convert such proof-of-concept into a public service.

5 Specific Actions with Reusable Outcomes in the Legal Domain

The proposed PMKI action meets the recommendations included in the European Interoperability Strategy (EIS[8]). The adherence to specific standards for describing language resources, and the creation of an interoperability platform to manage them, comply with the main approaches and "clusters" of the EIS (reusability of the solutions, interoperability service architecture in the EU multilingual context, implication of ICT on new EU legislation, as well as promotion of the awareness on the maturity level and of the shareability of the public administration services).

Similarly, the proposal meets the recommendations and principles of the European Interoperability Framework (EIF[9]), regarding multilingualism, accessibility, administrative simplification, transparency, and reusability of the solutions. The creation of a public multilingual knowledge infrastructure will allow EU public administrations to create services that can be accessible and shareable independently from the language actually used, as well as the SMEs to sell goods and service cross-border in a Digital Single Market.

As we have shown in section 2, the outcomes of such initiatives are prodrome for supporting document analysis, indexing and retrieval as well as cross-legislation access to legal content. In the next sections we will present the main actions foreseen in our contribution to the PMKI project and their potential in supporting the above objectives.

5.1 Comparative Study and Selection of Semantic Web Standards for Describing Multilingual Resources

A study has been conducted, embracing available web standards for multilingual resources at large, thus including multilinguality in ontologies, terminologies and specifically in lexical resources.

Due to the nature of the resources in PMKI, within the project different recommendations or even popular vocabularies will be adopted:

SKOS [22]: the W3C recommendation for formalizing thesauri, terminologies, controlled vocabularies and other knowledge resources characterized by shallow semantics. It is worth noticing that the terminological level solely supports the identification of concepts by giving them names (and alternative lexical references) but cannot be considered to be a lexicon nor any sort of advanced lexical resource, as any sort of lexical description taking into account phenomena such as morphology, lexical relations etc.. are considered, by definition, to be out of the scope of a thesaurus.

---

[8] http://ec.europa.eu/isa/documents/isa_annex_i_eis_en.pdf

[9] http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

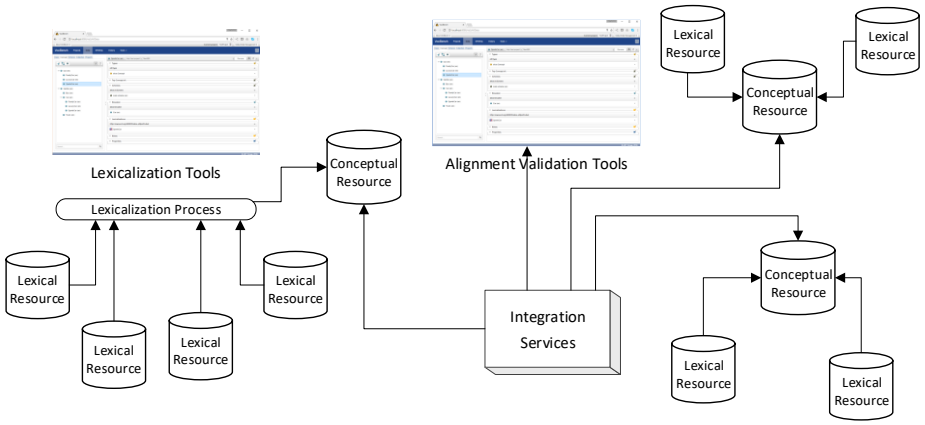

Fig. 2. PMKI Integration Framework, General Architecture

SKOS-XL [23]: As for the general definition of thesauri, SKOS does not address complex lexical descriptions of its elements. However, SKOS is extended by the SKOS-XL vocabulary which provides reified labels by means of the class skosxl:Label. SKOS terminological properties have their equivalents (identified by homonymous local names) in the new namespace, that is: skosxl:prefLabel, skosxl:hiddenLabel, skosxl:altLabel in order to relate concepts with these reified labels.

OntoLex-Lemon. We already described this model in section **Errore. L'origine riferimento non è stata trovata.**. As the model is relatively recent, there is still not much support for developing resources according to its vocabulary. As explained in the next section, we will develop a system, integrated into an already mature ontology/thesauri development environment, for the development of lexicons and for interfacing lexical knowledge with ontological one.

Other models and schemes. The above vocabularies represent the core of the selected models for development and alignment of resources in PMKI. The support for OntoLex will however not be limited to the enrichment of SKOS thesauri, and OWL ontologies or generic RDF datasets can be lexically enriched with OntoLex lexical descriptions with no loss of generality. Similarly, the above choices do not obviously prevent the adoption of specific metadata vocabularies, domain/application ontologies etc…

5.2 Systems for Semantic and Lexical Integration of Multilingual Resources

Support for integration will be implemented two-fold: by realizing a framework for alignment of semantic resources (thesauri, ontologies etc..) and by the development of a system for the development of lexicons according to the OntoLex vocabulary and for the lexical enrichment of semantic resources with lexical information.

Even though a pilot project in nature, PMKI is not a research project, it in fact aims at building up on well-established research results and existing technologies and at converging towards a concrete proposal for an integration framework.

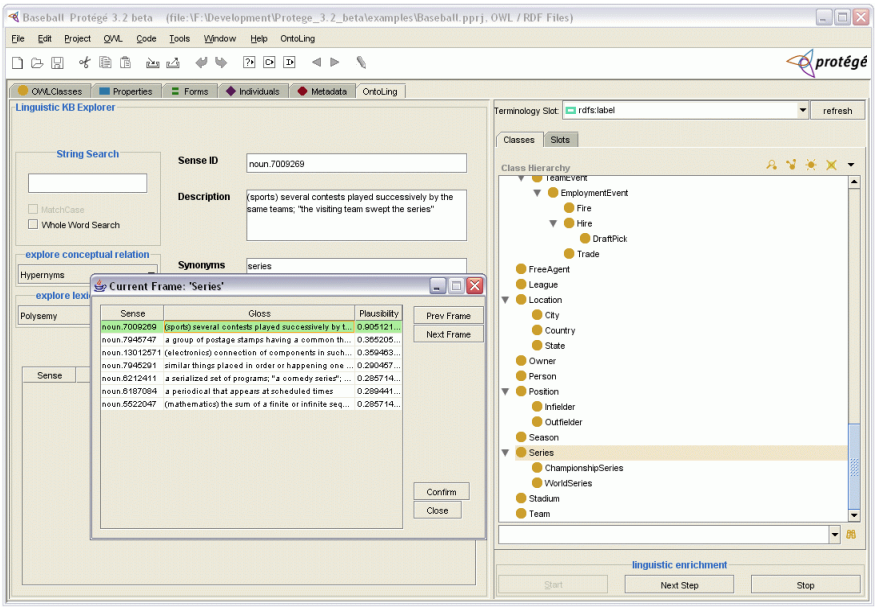

Fig. 3. A screenshot of OntoLing, a Protégé plugin for lexical enrichment of ontologies, that will be ported to VocBench and improved to exploit information from OntoLex resources

The general concept behind the framework is depicted in fig. 2, focused on the interaction of systems aimed at supporting the two tasks previously defined.

Semantic Integration Framework. The architecture foresees the presence of semantic integration services accessed by RDF management systems. The separation between the two is dictated by the different requirements in terms of interaction modalities, performance and results. RDF Management Services, whether single-user desktop applications or centralized collaborative platform, require high interaction with the user, averagely-low response times and, in the case of collaborative systems, the capacity to serve in real time several users accessing diverse projects. These platforms may offer manual or semi-automatic alignment functionalities, which though have to be performed with a low impact on system resource, and possibly replicated across several parallel requests. Conversely, Semantic Integration systems may instead act as token-based service providers, receiving requests to load and align datasets of considerable size, performing their function in non-trivial execution time due to the intensive analysis of the involved resources and dedicating considerable amount of resources to these tasks. After each alignment process has been completed, the alignment services may release the token to the requesting peer and start the next alignment task at the head of the request queue. A pool of processors may be considered in order to allow parallelization of alignment tasks.

The Semantic Integration System developed within the pilot project will be based on GENOMA [24], a, highly configurable alignment architecture, and on MAPLE [25], a metadata-driven component for the configuration of mediators, which will allow for seamless application of the same alignment techniques on datasets modeled according

to different modeling vocabularies, by providing vocabulary-specific implementations of the general analysis engine tasks. The manual/semi-automatic alignment capabilities will be provided by VocBench [26], a collaborative RDF management system for the development and maintenance of RDF ontologies thesauri and datasets, based on a service-oriented RDF management platform [27], recently updated to its third version [28] through another funded action of the ISA^2 program. As part of a coordinated action with the PMKI project, VocBench will also feature interaction modalities with the Semantic Integration system developed within PMKI.

Lexicon Development and Lexical Enrichment of Knowledge Resources. The OntoLex model is relatively young and, as such, it is still not widely supported by most mature technologies for data management. In a recent paper [29] describing the expressive power of VocBench 3 custom forms, the authors show how the custom form mechanism could be used to define complex lemon-patters. As VocBench 3 provides a general-purpose editing environment with specific facilities for the editing of SKOS and SKOSXL thesauri and OWL ontologies, extending the system with dedicated support for OntoLex-Lemon seems thus a natural way to cover this need.

In PMKI, VocBench will thus be improved to support the OntoLex-Lemon model in two different scenarios: developing Lexicons based on the OntoLex vocabularies and enriching semantic resources with lexical content. The two scenarios may be interwoven, as it will be possible to develop lexical entries specifically for semantic resources as well as reuse lexical content from existing lexicons in order to enrich the semantic resource with it. In most real applications, the two possibilities will not be alternative to each other: while a lexicon can usually provide domain-independent lexical entries, the description of specific concepts in a domain/application ontology often requires the definition of new complex terms, thus requiring in turn to state how the proper combination (at the lexical level) of single lexical entries would generate the accurate description of the conceptual elements. This implies the creation of further lexical entries describing the multiword, syntagmatic structure of the lexical representation of the complex concept. The inspiring work for such evolution of VocBench comes from past works (see Figure 3) concerning semi-automatic enrichment of ontologies by reuse of lexical resources [30] and exploitation of language metadata [31,32], which can now benefit from the standardization of this metadata for the Linked Open Data [33,34].

5.3 Realization of Concrete Semantic Alignments and Lexical Enrichments and Assessing

In the pilot project, a certain amount of alignments and lexicalizations will be produced. The objective is not only to produce the resources per-se, but to provide golden-standards that can be used in the later stage to evaluate the alignment systems.

Developing a gold standard mapping dataset is however not an easy task, due to the difficulty, even for humans, to lose potential matches in datasets of even modest size. Not incidentally, the Ontology Alignment Evaluation Initiative (OAEI), an initiative, implemented as a contest, which aims at evaluating the state of the art of ontology alignment tools, [35] mostly offers test beds of relatively small size and rarely updates

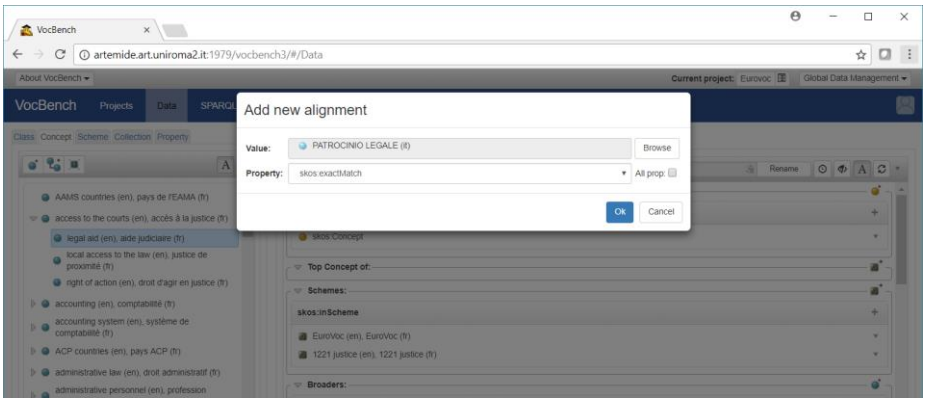

Fig. 4. aligning concepts between EU EuroVoc and Italian Senate's TESEO thesauri

the list of these mappings. Furthermore, evaluation of mappings sent by the participants, in the cases involving large datasets does not rely entirely on the standard and involves instead manual scrutiny, in order to take into account potentially correct mappings missing from the oracle.

We thus decided to divide the kind of contributions for dataset alignment in two steps: a vertical exploration of a humanly-computable subdomain of the two thesauri, and a larger attempt at mapping complete resources. The first result guarantees the creation of a reliably sound and complete set of mappings, while a larger alignment on the whole resources will be produced later on in the project, by means of semi-automatic processes, reusing the same systems that we will validate through the first result. An alignment that will be considered for production is the one – already mentioned in the example in section 2.1 – between EuroVoc and TESEO (see fig. 4).

Concerning lexicalizations, EuroVoc, as a central hub in the EU scenario, has, also in this case, been selected as the target conceptual resource. Candidate lexical resources for the results to be produced within the pilot are WordNet, being probably the most popular lexical resource and, for analogous reasons and more specifically in the EU scenario, IATE[10], the InterActive Terminology for Europe. However, both these resources do not provide the rich lexical and morphological descriptions representing the added value brought by the OntoLex model. For this reason, other resources such as BabelNet, or other lexicons still not modeled in OntoLex, will be taken into consideration. In the latter case the process will be two-fold: porting resources to OntoLex, which is a result per se, and then using them to lexicalize (part of) EuroVoc.

6 Conclusions

In this paper, we have presented the objectives and roadmap of the PMKI project and how its outcomes will directly and positively affect access to legal content and foster

[10] termcoord.eu/iate/

its exploitation in various scenarios. The multicultural, multi-jurisdictional and multilinguistic nature of the European Union has always been considered an asset rather than an obstacle, as it is through their differences that the Member States can learn from each other, benefiting from distinct experiences and approaches. Making these experiences truly and effectively comparable by lowering the language barriers and by harmonizing/connecting different though overlapping concepts and regulations is the objective of initiatives such as PMKI. Even though a pilot project, PMKI will aim to pave the way for analogous efforts while still contributing the community with tangible results in terms of systems and frameworks for alignment and lexicalization of heterogeneous resources.

References

1. Francesconi, E., Peruginelli, G.: Opening the legal literature portal to multi-lingual access. In : in Proceedings of the Dublin Core Conference, pp.37–44 (2004)
2. Antonini, A., Boella, G., Hulstijn, J., Humphreys, L.: Requirements of Legal Knowledge Management Systems to Aid Normative Reasoning in Specialist Domains. In Nakano, Y., Satoh, K., Bekki, D., eds. : New Frontiers in Artificial Intelligence. Lecture Notes in Computer Science. JSAI-isAI 2013. 8417. Springer, Cham (2013), pp.167-182
3. Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F.: Evaluation of ontolearn, a methodology for automatic population of domain ontologies. In : Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press (2005)
4. Pennacchiotti, M., Pantel, P.: Automatically Harvesting and Ontologizing Semantic Relations. In Buitelaar, P., Cimiano, P., eds. : Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Series: Frontiers in Artificial Intelligence. IOS Press (2008)
5. Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., eds.: Survey of the State of the Art in Human Language Technology. Cambridge University Press, Cambridge, UK (1997)
6. Calzolari, N., McNaught, J., Zampolli, A.: EAGLES Final Report: EAGLES Editors Introduction., Pisa, Italy (1996)
7. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. (1993)
8. Fellbaum, C.: WordNet: An Electronic Lexical Database. WordNet Pointers, MIT Press, Cambridge, MA (1998)
9. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998)
10. Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Marinelli, R., Magnini, B., Speranza, M., Zampolli, A.: ItalWordNet: A Large Semantic Database for the Automatic Treatment of the Italian Language. In : First International WordNet Conference, Mysore, India (January 2002)
11. Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M.: BALKANET: A Multilingual Semantic Network for the Balkan Languages. In : First International Wordnet Conference, Mysore, India, pp.12-14 (January 2002)

12. Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C.: Lexical Markup Framework (LMF). In : LREC2006, Genoa, Italy (2006)
13. Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon Model for Ontologies: Community Report, 10 May 2016. Community Report, W3C (2016) https://www.w3.org/2016/05/ontolex/.
14. Borin, L., Dannélls, D., Forsberg, M., McCrae, J.P.: Representing Swedish Lexical Resources in RDF with lemon. In : Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, pp.329-332 (2014)
15. Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J.P., Cimiano, P., Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In : Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., pp.401-408 (2014)
16. Eckle-Kohler, J., McCrae, J.P., Chiarcos, C.: lemonUby - a large, interlinked syntactically-rich lexical resources for ontologies. Semantic Web Journal (2015 (accepted))
17. Sérasset, G.: Dbnary: Wiktionary as a LMF based Multilingual RDF network. In : Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, pp.2466-2472 (2012)
18. Buitelaar, P.: Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. In Huang, C.-r., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prevot, L., eds. : Ontology and the Lexicon: A Natural Language Processing Perspective. Cambridge University Press (April 2010)
19. Cimiano, P., McCrae, J., Buitelaar, P., Montiel-Ponsoda, E.: On the Role of Senses in the Ontology-Lexicon. In Oltramari, A., Vossen, P., Qin, L., Hovy, E., eds. : New Trends of Research in Ontologies and Lexical Resources. Springer Berlin Heidelberg (2013), pp.43-62
20. Evans, V.: Lexical concepts, cognitive models and meaning-construction. Cognitive Linguistics 17(4), 491-534 (December 2006)
21. Chiarcos, C., McCrae, J., Cimiano, P., Fellbaum, C.: Towards Open Data for Linguistics: Linguistic Linked Data. In Oltramari, A., Vossen, P., Qin, L., Hovy, E., eds. : New Trends of Research in Ontologies and Lexical Resources. Springer Berlin / Heidelberg (2013), pp.7-25 10.1007/978-3-642-31782-8_2.
22. World Wide Web Consortium (W3C): SKOS Simple Knowledge Organization System Reference. In: World Wide Web Consortium (W3C). (Accessed August 18, 2009) Available at: http://www.w3.org/TR/skos-reference/
23. World Wide Web Consortium (W3C): SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). In: World Wide Web Consortium (W3C). (Accessed August 18, 2009) Available at: http://www.w3.org/TR/skos-reference/skos-xl.html
24. Enea, R., Pazienza, M.T., Turbati, A.: GENOMA: GENeric Ontology Matching Architecture. In Gavanelli, M., Lamma, E., Riguzzi, F., eds. : Lecture Notes in Computer Science. Proceedings of the IA*IA 2015 Conference. 9336. Springer International Publishing (2015), pp.303-315
25. Fiorelli, M., Pazienza, M.T., Stellato, A.: A Meta-data Driven Platform for Semi-automatic Configuration of Ontology Mediators. In Chair), N.C.(., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds. :

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 2014. European Language Resources Association (ELRA), Reykjavik, Iceland (2014), pp.4178-4183

26. Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., Keizer, J., Pazienza, M.T.: VocBench: a Web Application for Collaborative Development of Multilingual Thesauri. In Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A., eds. : The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science). Proceedings of the 12th Extended Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, 31 May - 4 June 2015. 9088. Springer International Publishing (2015), pp.38-53
27. Pazienza, M.T., Scarpato, N., Stellato, A., Turbati, A.: Semantic Turkey: A Browser-Integrated Environment for Knowledge Acquisition and Management. Semantic Web Journal 3(3), 279-292 (2012)
28. Stellato, A., Turbati, A., Fiorelli, M., Lorenzetti, T., Costetchi, E., Laaboudi, C., Van Gemert, W., Keizer, J.: Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond. In : 17th European Networked Knowledge Organization Systems (NKOS) Workshop (21st September 2017), Thessaloniki, Greece
29. Fiorelli, M., Lorenzetti, T., Pazienza, M.T., Stellato, A.: Assessing VocBench Custom Forms in Supporting Editing of Lemon Datasets. In : Language, Data, and Knowledge. Series: Lecture Notes in Computer Science. First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings. 10318. Springer Cham, Galway, Ireland (2017), pp.237-252
30. Pazienza, M.T., Stellato, A.: An Environment for Semi-automatic Annotation of Ontological Knowledge with Linguistic Content. In Sure, Y., Domingue, J., eds. : The Semantic Web: Research and Applications (Lecture Notes in Computer Science). 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11-14, 2006, Proceedings. 4011. Springer (2006), pp.442-456
31. Pazienza, M.T., Sguera, S., Stellato, A.: Let's talk about our "being": A linguistic-based ontology framework for coordinating agents. Applied Ontology, special issue on Formal Ontologies for Communicating Agents 2(3-4), 305-332 (December 2007)
32. Pazienza, M.T., Stellato, A., Turbati, A.: Linguistic Watermark 3.0: an RDF framework and a software library for bridging language and ontologies in the Semantic Web. In : 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008), Rome, Italy, December 15-17, 2008, CEUR Workshop Proceedings, FAO-UN, Rome, Italy, vol. 426, p.11 (2008)
33. Fiorelli, M., Stellato, A., Mccrae, J.P., Cimiano, P., Pazienza, M.T.: LIME: the Metadata Module for OntoLex. In Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A., eds. : The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science). Proceedings of the 12th Extended Semantic Web Conference (ESWC 2015), Portoroz, Slovenia, May 31 - 4 June, 2015. 9088. Springer International Publishing (2015), pp.321-336
34. Fiorelli, M., Pazienza, M.T., Stellato, A.: An API for OntoLex LIME datasets. In : OntoLex-2017 1st Workshop on the OntoLex Model (co-located with LDK-2017), Galway (2017)
35. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158-176 (January 2013)