Collaborative Maintenance of EDOAL Alignments in VocBench

Armando Stellato[0000-0001-5374-2807], Manuel Fiorelli[0000-0001-7079-8941], Tiziano Lorenzetti[0000-0001-5676-8877], Andrea Turbati[0000-0002-6214-4099]

University of Rome "Tor Vergata", Department of Enterprise Engineering, Via del Politecnico 1, 00133 Rome, Italy
stellato@uniroma2.it, fiorelli@info.uniroma2.it, tiziano.lorenzetti@gmail.com, turbati@info.uniroma2.it

Abstract. Ontology Alignment, intended in its broadest meaning of alignment between datasets of different nature – thesauri, ontologies, and even mere instance data – is a well-known practice aiming at realizing semantic links between datasets on the (Semantic) Web. Considerable investigation has been carried on the automatic computation of alignments and on how to assess the quality of such process. This is indeed a critical aspect, considering the non-trivial size of many datasets. However, since human intervention is in any case essential, no less care should be paid on scalability both in terms of distribution of work and of maintenance of achieved results within the lifecycle of the aligned resources. In this paper we guide the reader through the diverse solutions that have been implemented in VocBench, a collaborative editing platform for RDF datasets, under a holistic approach to collaborative alignment development and maintenance.

1 Introduction

The Semantic Web [1] has offered a powerful stack of standard languages and protocols for modeling, sharing and reuse of knowledge on the Web. However, the advantages brought by metadata standards and infrastructures for shareability of information cannot avoid (but can support) the reconciliation work needed on the information content. Different, domain-overlapping, redundant to different extents, ontologies, thesauri, vocabularies, datasets etc. are expected to emerge on the Web to satisfy specific needs and exigencies and, at different points in time, are expected as well to be – somehow – "reconciled" out of their heterogeneities, for interoperability's sake.

This "reconciliation" takes the form of alignments, that is, sets of correspondences between the different entities that populate lexical and semantic resources on the Web. The expression "ontology alignment" is often used in a broader sense than the one that the first word of the term would suggest. "Ontology" is in this case a synecdoche for ontologies, thesauri, lexicons and any sorts of knowledge resources modeled according to core knowledge modeling languages for the Semantic Web, which were shared and made available on the Web itself. The expression ontology alignment thus defines the task of discovering and assessing alignments between ontologies and other data models of the RDF family; alternative expressions are *ontology mapping* or *ontology matching* (as the produced alignments are also referred to as *matches*). In the RDF jargon, and following the terminology adopted in the VoID metadata vocabulary [2], a set of alignments is also called a *Linkset*.

The production of alignments is an intensive and error prone-task; for this reason, several approaches for automating the task have been devised [3] since the early years of the Semantic Web. An Ontology Alignment Evaluation Initiative[1] [4] is also held every year since 2004 with the intent of evaluating available tools against benchmarks consisting in well-assessed alignments between notable semantic resources (mostly ontologies, and some thesauri). The task are also divided into T-Box/Schema matching, dealing – as the name suggests – with the alignment of ontology vocabularies (i.e. world models including classes and properties), and instance matching, involving the creation of links between domain objects represented in different datasets.

In this work we present our approach for holistic management of alignments within our collaboration platform called VocBench [5]. The proposed solution is characterized by a multitude of tools, features and functionalities, all interconnected within a more general platform for collaborative management of RDF vocabularies and datasets.

2 Related Work

Even if shadowed for the most by research on automatic computation of alignments, the importance of alignment maintenance and of collaboration on their evolution has not been underestimated.

Limiting our overview to ontologies (as there is a large literature on database schema matching), one of the first, if not the first, collaborative platforms for alignment maintenance is probably Chimaera [6], a web based tool for merging ontologies and checking the correctness of ontologies. Chimera mixes automatic procedures (integrated with the platform) with interfaces for displaying the knowledge content and allowing for users to validate the produced mappings.

Another relevant work is represented by the MAFRA toolkit [7], a mapping framework for distributed ontologies that introduced the concept of transformation. In MAFRA, the objective is not just to find correspondences, rather being able to transform content objects that are specified in one dataset into their corresponding version (which may vary in the structure) on the aligned dataset. MAFRA is antecedent to the release of SPARQL as a query and update language for the Semantic Web, hence the need for expressing these transformations through a dedicated language.

The Alignment API [8] provided several contributions to the community, as other than representing, as their name suggest, a reusable set of API for managing alignments, they come together with an Alignment Server – an online service for alignment manipulation (creation, modification, display) – and a dedicated language, EDOAL, which has become a standard de-facto interchange format for representing alignments as first-class citizens i.e. they are not just triples using alignment predicates from OWL and

[1] http://oaei.ontologymatching.org/

SKOS to map resources from a source and a target dataset, rather resources themselves containing links and providing further information such as the strength of the alignment. Thanks to the reification of the alignments, it is possible to attach to them further metadata. VocBench adopts EDOAL and indeed extends it with a vocabulary of further information for both storing the progress in the validation of alignments and, conversely, exploiting the data for evaluating the quality of automatic alignment tools.

More recently, the Agreement Maker [9] platform capitalized the experiences of past approaches, specifically in the geospatial domain, with a visual software tool for creating mappings between two ontologies, generating an agreement document.

Nowadays, support, at different levels, for collaborative alignments is also provided by most collaborative editing platforms, such as Semantic Web Company's PoolParty [10] and TopQuadrant's TopBraid Enterprise Data Governance (EDG)[2]

3 The VocBench Collaborative Editing Platform

VocBench 3 is a free and open-source advanced collaboration environment for creating and maintaining ontologies, thesauri, code lists, authority tables, lexicons, link sets and datasets, in compliance with Semantic Web standards recommended by the W3C.

VocBench 3 (from now on, VocBench or simply VB) is used to maintain vocabularies and ontologies in a wide number of domains, including the agrifood sector (e.g. [11,12], see also [13] for a more thorough overview), government data, thesauri, ontologies (e.g. the Teseo thesaurus[3] of the Italian Senate of the Republic or EU's multilingual thesaurus EuroVoc[4] and many other semantic resources managed by the European Commission, such as the Common Metadata Model [14] ontology, thus exploiting the widened support of VB3 for OWL ontologies) and several others.

The user community's main point of interaction and support is the VocBench discussion group[5]. The group (as of 20th January 2021) counts 213 members.

The VocBench 3 project is funded by Action 1.1 of the ISA2 Programme of the European Commission for "Interoperability solutions for public administrations, businesses and citizens". The VocBench site[6] contains documentation, download links and other references. VocBench supports different user profiles in fulfilling their job by providing different functionalities tailored for their roles. As a software suite, many of the tools offered by VocBench are available as separate components that can be integrated in other software. Depending on the standard being adopted and the type of user, the system provides several user-tailored facilities for easily modeling the needed resource and toggles or makes optional (with properly conceived default settings) various features. Quality checking is provided by dedicated Integrated Constraint Validation tools (ICVs) and by support for SHACL shapes.

[2] https://www.topquadrant.com/products/topbraid-enterprise-data-governance/
[3] http://www.senato.it/tesauro/teseo.html
[4] https://data.europa.eu/euodp/data/dataset/eurovoc
[5] http://groups.google.com/group/vocbench-user
[6] http://vocbench.uniroma2.it

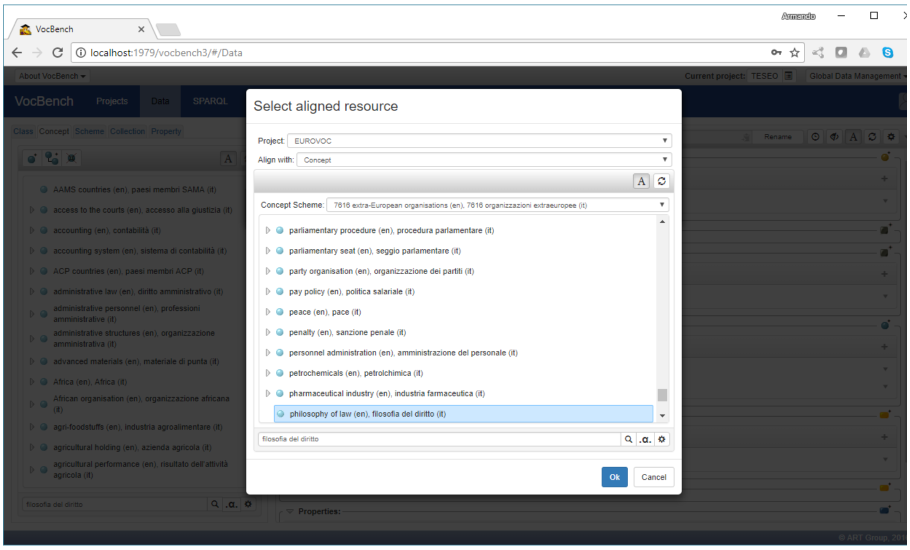

Figure 1. on-the-fly manual alignment of concepts from the TESEO thesaurus to concepts from EuroVoc

4 Alignment Support in VocBench at Large

VocBench does not provide a single entry point for addressing alignment development, rather alignment support is a pervasive aspect appearing in diverse points of the user experience. Roughly, we can identify the following entry points for performing/maintaining alignments:

- Alignment from within the resource-view
  - Manual alignment (search based)
  - Semi-automatic (search keywords based on available labels)
- Alignment Validation
  - Input coming from a static EDOAL file
  - Input coming from the invocation of an automatic alignment system
- EDOAL projects

4.1 Alignment from within the Resource View

The first possibility represents the quickest and most flexible choice, as it is embodied by an action that can be performed from within any kind of project (i.e. for any nature of the dataset, be it an ontology, thesaurus, lexicon or mere data), linking to both internally managed datasets (i.e. other projects within the same VocBench installation) and datasets on the Web (by exploiting well known practices for discovering them and exploiting available metadata for knowing how to inspect the content and what to look for), without any prior preparation/setup. From within the resource view, the panel that

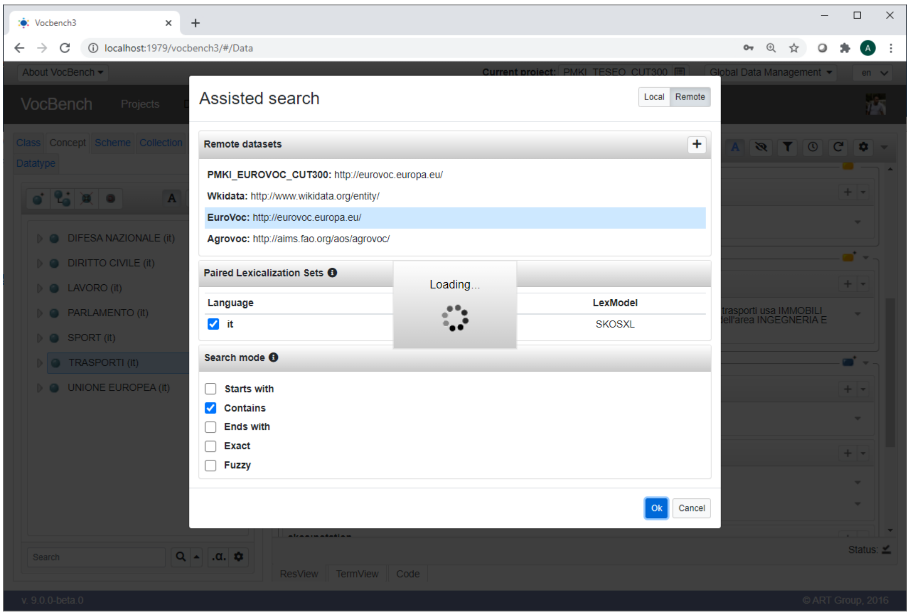

Figure 2. alignment through assisted search of concepts from the TESEO thesaurus to concepts from EuroVoc directly from the SPARQL endpoint of the original dataset on the Web

shows all the details of a selected resource, it is possible to select an option for aligning it to an external resource from another dataset, then the user is prompted for three possible choices:

- Aligning to a resource from a project, by browsing the content of the project
- Aligning by means of assisted search; this option is available for both projects and external datasets hosted on the Web
- Manually enter the value of an aligned resource

In the first option, a list of projects generated by filtering those that can be accessed by the project being managed (depending on the ACL, see later) is shown to the user: the selected one will be the "target project" for choosing the aligned resource. After the user selects the target project, a dialog such as the one in Figure 1 is prompted to the user, who can then explore the resources through convenient browsing panels (i.e. class/property/concept/collection trees for owl classes, SKOS concepts/collections or RDF properties in general, lists for concept schemes and class tree and related lists of instances for each class in the case of owl individuals). The ACL (Access Control List) is an access matrix in VocBench telling which projects can access the content of other projects. This access is indeed a delegation to users, so that if project A grants access to project B, this means that any user that has been granted access to project B, no matter their relationship with project A, may access the content of this latter by means of the permission from A to B and the delegation of all users of B.

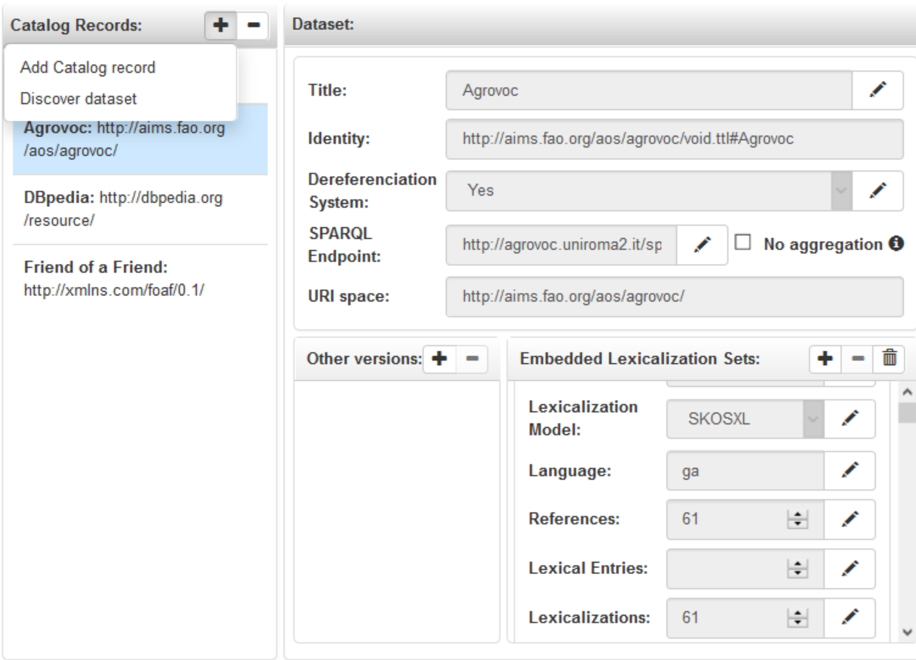

Figure 3. User interface of the Metadata Repository inside VocBench 3

The second option for alignment exploits dataset's metadata in order to automatically select the search keywords for the target dataset. The keywords basically come from the labels of the resource to be aligned (the one from the source dataset). What is interesting is the choice of the language of the labels (Figure 2). To reduce the complexity of the query, which would imply (in case the source is characterized by lexicalizations in many languages) an OR clause of several terms, the system presents the lexicalization languages that are common to both the datasets and ranks them according to an average of the quality of their related lexicalizations for both datasets. This quality is mainly measured in terms of coverage of the dataset (the percentage of resources that are lexicalized in that given language) and, on a second order, on the lexical richness in that language (the overall number of lexicalizations, thus revealing the presence of alternative expressions, synonyms etc..).

Finally, dealing with the trivial case of already assessed alignments only needing to be added to a dataset, it is possible to manually add a link to a resource by explicitly expressing that resource through its IRI.

4.2 The Metadata Registry

The way the lexicalization metadata (used to power the semi-automatic alignment) is obtained is another peculiar characteristic of the VocBench platform, which tries to exploit Semantic Web principles to their full potential. VocBench features, among its diverse components, a *metadata registry* [15]. This registry contains metadata related

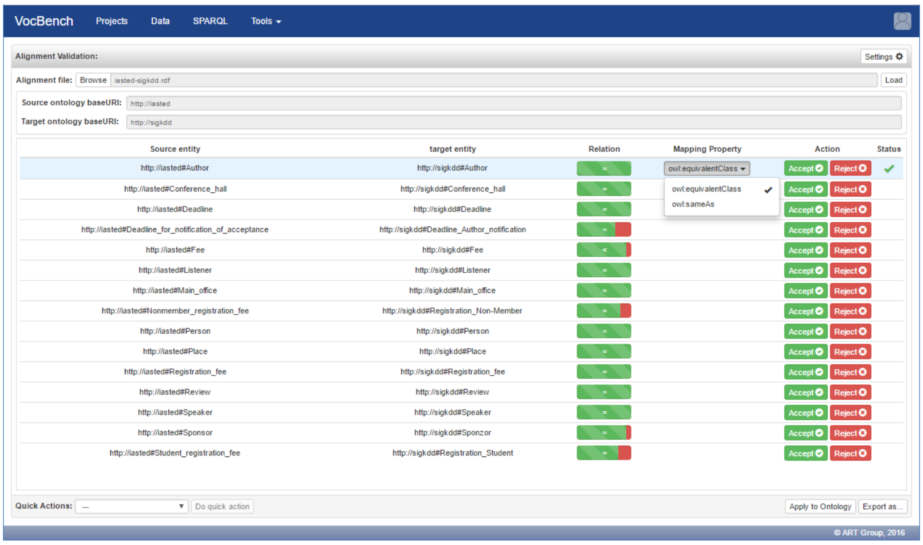

Figure 4. The Alignment Validation interface of VocBench 3

to both the datasets edited within the platform and other datasets on the Web, and supports different strategies to acquire it.

The dataset metadata's model is a combination of different vocabularies (among these: DCAT [16], VoID [2] and LIME [17]), enriched with some system-specific information. In particular, LIME (LInguistic MEtadata vocabulary), developed by these same authors in the context of the OntoLex community group, and based on previous research on automatic coordination for alignments [18,19], provides rich statistical and qualitative information about the lexical asset of datasets and drives the coordination of resources and algorithms for performing alignments between heterogeneous sources.

The description of local projects combines information found in some configuration files with metadata generated through a profiler provided by our LIME API [20], while the description of remote projects may come directly from their metadata files exposed on the web or can be profiled as well (at least for the statistical part). A discovery mechanism allows for finding metadata files on the Web starting from just the IRI of any single resource belonging to the dataset, by following well known principles for publishing datasets[7]. An editor for the registry (Figure 3) allows for the modification of the metadata gathered through the above different way and for grouping diverse datasets as different versions and/or distributions of a same one.

4.3 Alignment Validation

VocBench features an Alignment Validation page, providing several instruments (indeed, due to the later evolutions, this page is soon to be renamed "Alignment Management") for building and validating alignments.

[7] `http://linkeddatabook.com/editions/1.0/`

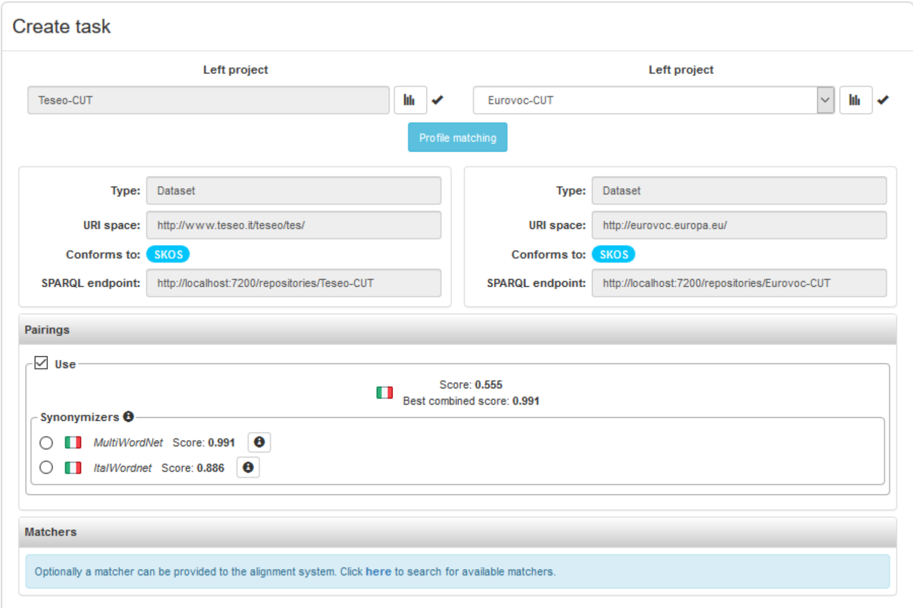

Figure 5. Defining an alignment scenario through the MAPLE UI of VB3

Once the page has been accessed, the first option being prompted is whether to load an already available alignment document – i.e. a set of mappings between two datasets expressed according to the EDOAL [8] vocabulary – or to produce one by means of an external alignment system.

In the first case, a set of alignments (also referred as *linkset*) to another dataset is loaded in the validation interface (see Figure 4). The alignment interface shows a list of aligned resources. For each alignment, a resource from the source dataset (the one being managed within the active project) and one from the target dataset are linked to each other. A "relation" expressed, according to the EDOAL vocabulary, the nature of this link. EDOAL provides an (extensible) set of relations providing shallow semantics for the concept of broader/narrower/equivalent. The "strength" of the relation is expressed by a quantity (in the range of 0..1) representing how reliable the link is. Each link can be accepted or rejected; in the former case, it is possible to define a mapping property from the OWL or SKOS vocabulary that will instantiate the relation. We have extended the EDOAL standard by allowing for storing both the progress in the validation and the mapping property that has been selected. The use of stored alignments is twofold: it can be simply used to restore an incomplete validation process in order to finish the job or, even when completed, could be used as an oracle for evaluating the quality of automatically produced alignments.

4.4 External Alignment Systems

A second option for alignment validation enables communication with external alignment systems in order to produce fresh new alignments to be validated. This

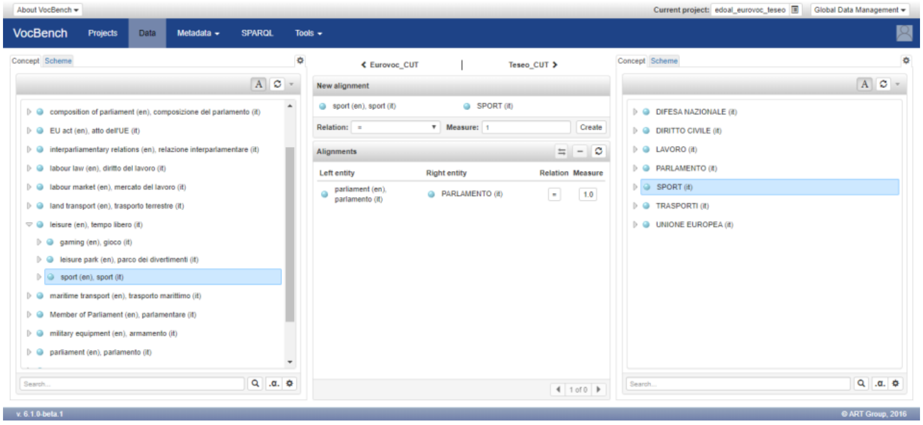

Figure 6. Maintaining alignments between two datasets in an EDOAL project

communication is managed by a dedicated component in VocBench called MAPLE. Given a matching task, MAPLE first analyzes the context of the alignment by inspecting the metadata of the two datasets (if the metadata is missing, dedicated profilers can produce this metadata on the fly and store it in the metadata registry of VocBench) and establishing the best matching lexicalizations from the two involved datasets, the requirements for their analysis (e.g. being able to read a certain lexical model, such as SKOS-XL or OntoLex) and other boundary conditions. MAPLE can also produce information about known lexical datasets (e.g. wordnets) that can be used as supporting resources for expanding the available lexical knowledge and provide a better bridge between the lexical description of the two datasets to be aligned. The result of this analysis is then compiled into a task report that is communicated to a downstream matching system, which executes the matching process. It is worth noticing that, as such, the system has no specific performance but completely demands the computation of the alignments to the external, connected, alignment systems.

The principle of communication with MAPLE is realized through an open API for alignment[8] [21] that external systems may comply with. Currently, there are two systems adopting this API: Genoma [22], a simple yet highly configurable alignment system developed by part of the same team that developed VocBench, and NAISC[9] [23], an alignment tool created by the SFI Insight Centre for Data Analytics in the context of the Horizon 2020 ELEXIS project (grant agreement No 731015).

4.5 EDOAL Projects

The last (also in temporal terms, as it has been introduced in version 7 of VocBench, released on the first quarter of 2020) addition to VocBench for supporting development of alignments is the introduction of EDOAL projects. EDOAL projects have alignment

8 https://bitbucket.org/art-uniroma2/maple/src/master/maple-alignment-services-api/src/main/openapi/alignment-services.yaml

9 https://elex.is/new-out-now-naisc-1-0/

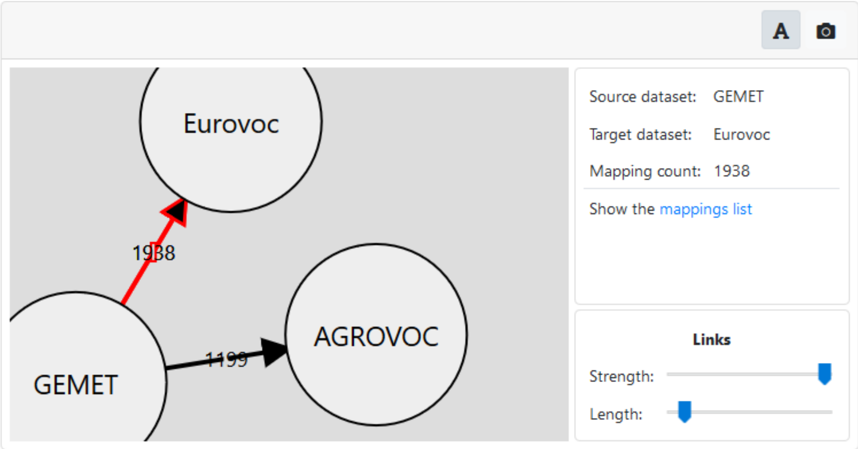

Figure 7. Browsing linksets in PMKI

as their first-class citizens. The typical scenario for this kind of project is collaborative development of alignments, independent from the maintenance of the aligned datasets. For instance, two organizations maintaining their own datasets may decide to collaborate on the development of a common linkset among their datasets. Obviously, neither of the two organizations want to give the other access to their maintained copy of the dataset, they thus setup an independent project, linked in read-only mode to a copy of the (or even the originally maintained) datasets. There data can be created, evolved, in an alignment-centric way: there is no source/target dataset; alignments are reified and equally maintained by both parties and observed by comparing the two datasets side-by-side (Figure 6), who can then export them as simple mapping triples, placing their own resources as subject.

5 Hosting an Open Data Portal [PMKI]

PMKI is a project funded by the ISA2 programme of the EU, aimed at the development of open data portals focusing on terminological and linguistic content. The software powering the portal is a sort of read-only VocBench, including the resource view and much of the browsing views, with a focus on efficiency and streamlined fruition of content. Besides the browsing-oriented user experience, the system features dedicated capabilities mostly based on the possibility to explore all of the hosted datasets as a single, large, resource, while still keeping provenance information about which data comes from which project. These features include, among others, a global free-text search over all datasets and machine translation API (benefiting from cross-lingual links provided by multilingual lexical resources). Another relevant feature lies in the possibility to browse linksets between datasets through a dedicated graph (see Figure 7) exploration where each node represents a dataset as a whole and the arcs represent the linksets. For each linkset, it is then possible to list its mapping statements and browse the involved aligned resources.

6 Conclusions

In this paper we have detailed, through the example of the VocBench platform, several supporting features that we consider of primary importance for the maintenance of alignments during their whole lifetime. The described features have been developed through a few years, across different iterations of the platform, however they all were part of a single design that aims at conceiving alignments as fundamental building blocks of the Semantic Web architecture, needing not only to be discovered and developed, but also maintained in a continuous lifecycle, kept alive by the same resources they link. Future work will focus even more on this latter aspect, improving the coordination between projects in order to keep alignments up-to-date with the evolution of the aligned datasets. There are still some functionalities for reporting broken links, but we plan to go further in exploring and exploiting the available alignments and the aligned content in order to automatically provide suggested actions to perform.

References

1. Berners-Lee, T., Hendler, J.A., Lassila, O.: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 279(5), 34-43 (2001)
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary (W3C Interest Group Note). In: World Wide Web Consortium (W3C). (Accessed March 3, 2011) Available at: http://www.w3.org/TR/void/
3. Euzenat, J., Shvaiko, P.: Ontology Matching 2nd edn. Springer-Verlag Berlin Heidelberg (2013)
4. Dragisic, Z., Eckert, K., Euzenat, J., Faria, D., Ferrara, A., Granada, R., Ivanova, V., Jimenez-Ruiz, E., Kempf, A.O., Lambrix, P., Montanelli, S., Paulheim, H., Ritze, D., Shvaiko, P., Solimando, A., Trojahn, C., Zamazal, O., Cuenca Grau, B.: Results of the ontology alignment evaluation initiative 2014. In : 9th International Workshop on Ontology Matching, October 20, 2014, Riva del Garda, Trentino, Italy, vol. 1317, pp.61-104 (2014)
5. Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., Keizer, J.: VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. Semantic Web 11(5), 855-881 (Jan 2020)
6. McGuinness, D.L., Fikes, R., Rice, J., Wilder, S.: The chimaera ontology environment. In : In Proceedings of AAAI, pp.1123-1124 (2000)
7. Maedche, A., Motik, B., Silva, N., Volz, R.: Mafra — a mapping framework for distributed ontologies. In : International Conference on Knowledge Engineering and Knowledge Management, pp.235-250 (2002)
8. David, J., Euzenat, J., Scharffe, F., Trojahn dos Santos, C.: The Alignment API 4.0. Semantic Web Journal 2(1), 3-10 (2011)
9. Cruz, I.F., Sunna, W., Makar, N., Bathala, S.: A visual tool for ontology alignment to enable geospatial interoperability. Journal of Visual Languages & Computing 18(3), 230-254 (June 2007)
10. Schandl, T., Blumauer, A.: PoolParty: SKOS Thesaurus Management Utilizing Linked Data. In Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral,

L., Tudorache, T., eds. : The Semantic Web: Research and Applications (Lecture Notes in Computer Science) 6089. Springer, Berlin, Heidelberg (2010), pp.421-425

11. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The AGROVOC Linked Dataset. Semantic Web Journal 4(3), 341–348 (2013)
12. Anibaldi, S., Jaques, Y., Celli, F., Stellato, A., Keizer, J.: Migrating bibliographic datasets to the Semantic Web: The AGRIS case. Semantic Web (2015) ISSN for Online version: 2210-4968.
13. Caracciolo, C., Aubin, S., Jonquet, C., Amdouni, E., David, R., Garcia, L., Whitehead, B., Roussey, C., Stellato, A., Villa, F.: 39 Hints to Facilitate the Use of Semantics for Data on Agriculture and Nutrition. Data Science Journal 19(1), 47 (2020)
14. Francesconi, E., Küster, M.W., Gratz, P., Thelen, S.: The Ontology-based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. In : International Conference on Electronic Government and the Information Systems Perspective (EGOVIS 2015), Valencia, Spain (2015)
15. Fiorelli, M., Stellato, A., Lorenzetti, T., Schmitz, P., Francesconi, E., Hajlaoui, N., Batouche, B.: Metadata-driven Semantic Coordination. In Garoufallou, E., Fallucchi, F., William De Luca, E., eds. : Metadata and Semantic Research (Communications in Computer and Information Science). 13th International Conference, MTSR 2019, Rome, Italy, October 28–31, 2019, Revised Selected Papers. 1057. Springer, Cham (2019)
16. World Wide Web Consortium (W3C): Data Catalog Vocabulary (DCAT). In: World Wide Web Consortium (W3C). (Accessed January 16, 2014) Available at: http://www.w3.org/TR/vocab-dcat/
17. 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
18. 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)
19. 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)
20. 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)
21. Fiorelli, M., Stellato, A.: A Lime-Flavored RESTAPI for Alignment Services. In Ionov, M., McCrae, J.P., Chiarcos, C., Declerck, T., Bosque-Gil, J., Gracia, J., eds. : Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020). European Language Resources Association (2020), pp.52-60
22. 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
23. McCrae, J.P., Buitelaar, P.: Linking Datasets Using Semantic Textual Similarity. Cybernetics and Information Technologies 8(1), 109-123 (2018)