One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web

John P. McCrae[1], Penny Labropoulou[3], Jorge Gracia[2], Marta Villegas[4], Víctor Rodríguez Doncel[2], and Philipp Cimiano[1]

[1] Cognitive Interaction Technology, Excellence Cluster, Bielefeld University, Germany
{cimiano, jmccrae}@cit-ec.uni-bielefeld.de
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{jgracia, vrodriguez}@fi.upm.es
[3] ILSP/Athena R.C., Athens, Greece
penny@ilsp.athena-innovation.gr
[4] University Pompeu Fabra, Barcelona, Spain
marta.villegas@upf.edu

Abstract. **Keywords:** keywords

1 Introduction

The study of language and the development of natural language processing (NLP) applications requires the access to language resources (LRs). Recently, several digital repositories that index metadata for LRs have emerged, supporting the discovery and reuse of LRs. One of the most remarkable of such initiatives is META-SHARE [12] (www.meta-share.eu), an open, integrated, secure and interoperable exchange infrastructure where LRs are documented, uploaded, stored, catalogued and announced, downloaded, exchanged and discussed, aiming to support reuse of LRs. Towards this end, META-SHARE has developed a rich metadata schema that allows to describe aspects of LRs accounting for their whole lifecycle from their production to their usage. The schema has been implemented as an XML Schema Definition (XSD). Descriptions of specific LRs are available as XML documents. Yet, META-SHARE is not the only metadata repository for language resources; other repositories include the CLARIN Virtual Language Repository[5] [2] as well as the LRE-Map[6] [5]. The metadata schemes of these different repositories vary with respect to their coverage and the set of specific metadata captured. All these repositories are complementary and index different language resources PL: rephrase: their sources are different but they have overlapping sets of LRs. Currently, it is not possible to query all these repositories in an integrated and uniform fashion. We argue that the Web of Data is a natural scenario for exposing LRs metadata in order to allow their automated discovery, share and reuse by humans or software agents.

[5] http://catalog.clarin.eu/vlo/?1
[6] http://www.resourcebook.eu/searchll.php

In this paper we contribute to the interoperability of all these repositories by developing an ontology in the Web Ontology Language (OWL) [11] that allows to represent the metadata schemes of these repositories uniformly, thus achieving an important first crucial step to establish interoperability between these repositories. OWL allows for a higher expressive level than the original XML representation as well as the application of semantic reasoning techniques (i.e., to infer new knowledge that were not initially declared). Also, the XML-based representation has proved inefficient when linking metadata resources. In fact, the use of RDF for representing the LRs metadata underlying the OWL ontology enables direct mechanisms for linking between metadata of different LRs and between metadata of LRs and other external sources. The resulting data is lighter, better suited for exploitation and eases further extensions and links with external resources (e.g., DBpedia). Finally, the use of Semantic Web techniques enable also standardized means of accessing the data (e.g., via SPARQL) thus not relying on domain-specific data formats or proprietary APIs. The proposed ontology is based on the ontology developed by Villegas et al. [15] for the UPF's META-SHARE node, covering part of the original schema, however extending this initial effort to the whole schema and all LRs and incorporating the consensus reached in the context of the W3C Linked Data for Language Technologies (LD4LT) Community Group[7]. As a proof of concept of this ontology, we describe how we have mapped metadata records from the above mentioned three repositories (META-SHARE, CLARIN, LRE-Map) into this ontology. Further, we describe *LingHub* [8], a portal that indexes and provides access to all these metadata records from the mentioned repositories. Our approach has several advantages. Firstly, the use of Semantic Web techniques enables standardized means of representing, linking, and accessing the data. Secondly, we hope that the use of this ontology will enable the representation of metadata in a manner that allows existing resources to adopt a common core vocabulary, while still being able to represent specific extensions to their existing model and we evaluate this hypothesis by reference to the CLARIN and LRE-Map data models. The remainder of this paper is structured as follows: in section 2 we will describe the related work in the fields of LR metadata and metadata harmonization. The development of the META-SHARE ontology is described in section 3 as well as its conversion to RDF and how the ontology was used for other data sources in that resource. Finally, in section 5 we consider the broader impact of this ontology as a tool for computational linguists and as a method to realize an architecture of (linked) data-aware services.

2 Related Work

The task of finding common vocabularies for linguistics is of wide interest and several general ontologies for linguistics have been proposed. The General Ontology for Linguistic Description [7, GOLD] was proposed as a common model

[7] https://www.w3.org/community/ld4lt

[8] http://linghub.org/

for linguistic data, but its relatively limited scope and low coherence has not lead to wide-spread adoption. An alternative approach that has been proposed is to use ontologies to create coherence among the resources, in particular either by using ontologies to align different linguistic schemas [6] or by means of agreed identifiers [10]. As regards LRs, there are as many metadata schemas for their descriptions as catalogs and repositories for their presentation (e.g. ELRA, LDC, OLAC) and communities describing them (e.g. TEI for humanities scholars, CES for the language technology domain, etc.). The most widely accepted schema is the one suggested by Open Language Archives Community [1, OLAC] which builds on the Dublin Core metadata but which has been criticised as too minimal. Extending the principle of linking concepts through identifiers (stored in the ISOcat Data Category Registry), the Component Metadata Infrastructure [4] suggested and maintained by CLARIN, attempts to bring together "components", which consist of semantically close elements, in order to be shared among different communities when producing "profiles" for specific LR types; however, this has not been achieved (cf. 3.5) and the VLO resorts to ISOcat links for aggregating similar resources. However, as we observe in section 3.5, this has in practice merely resulted in each contributing institute using its own scheme, with very little commonality between different institutes. To improve this situation it was recently proposed that the conversion of these CMDI schemas to RDF would enable better interoperability [14], however it is not clear if this project has been realized.[9] Other initiatives aiming to bring together LRs include, among others, datahub [10] which targets datasets described by LR providers, the DiRT Directory [11] and TERESAH [12] focusing on tools for scholars and the Linguistic Linked Open Data [13] collecting descriptions of LRs available in Linked Data format.

3 The META-SHARE OWL Ontology

3.1 Original MS XSD schema[PL]

The design of the META-SHARE schema [8] has been based upon previous similar efforts and metadata schemas used for the description of LRs as well as user needs recorded for the META-SHARE infrastructure. It has been designed not only as an aid for LRs' search and retrieval processes but also as a means to fostering their production, use and re-use by bringing together knowledge about LRs and related objects and processes, thus encoding information about the whole lifecycle of the LR from production to usage stages. The central entity

[9] JPM: I emailed Menzo Windhouwer about this and may change this statement based on his response, if any; PL: in lrec they said they have rdfized the schemas but there has been no implementation as far as I know; take out unless we're sure
[10] http://datahub.io/
[11] http://dirtdirectory.org/
[12] http://staging.teresah.php.dev.dasish.eu/
[13] http://linguistics.okfn.org/resources/llod/

of the META-SHARE schema is the LR *per se*, which encompasses both **data sets** (e.g., textual, audio and multimodal/multimedia corpora, lexical data, ontologies, terminologies, computational grammars, language models) and **technologies (tools/services)** used for their processing. In addition to the central entity, other entities are also documented in the schema; these are reference documents related to the LR (papers, reports, manuals etc.), persons/organizations involved in its creation and use (creators, distributors etc.), related projects and activities (funding projects, activities of usage etc.), accompanying licenses, etc., all described with metadata taken as far as possible from relevant schemas and guidelines (e.g. BibTex for bibliographical references). PL: figure here? The META-SHARE schema proposes a set of elements to encode specific descriptive features of each of these entities and relations holding between them, taking as a starting point the LR. Following the CMDI approach, these elements are grouped together into "components". The core of the schema is the `resourceInfo` component (Figure 1– JPM where is this??), which subsumes

- administrative components relevant to all LRs, e.g. `identificationInfo` (name, description and identifiers), `distributionInfo` (licensing and IPR information), `usageInfo` (information about the intended and actual use of the LR) etc.
- components specific to the resourceType (corpus, lexical/conceptual resource, language model, tool/service) and mediaType (text, audio, video, image) combinations of the LR cater for the encoding of information relevant to text, audio, etc. parts of corpora, lexical/conceptual resources, etc. (e.g. language, formatting, classification).

The META-SHARE schema has been implemented as an XSD (available at GITHUB). An integrated environment supports the description of LRs, either from scratch or through uploading of XML files adhering to the META-SHARE metadata schema, as well as browsing, searching and viewing of the LRs.

3.2 Formal modelling and mapping issues [MV, JPM, PL]

PL: I think this paragraph should be turned into bullets in the list our decisions of mapping, one for the mapping of components/elements and one for the simplification rule In the META-SHARE XSD schema, *elements* are formalized as simple elements whereas *components* are formalized as complex-type elements. When mapping the XSD schema to RDF, *elements* can be naturally understood as properties (e.g. name, gender, etc.). *Components* (i.e. complex-type elements), however, deserve a careful analysis. General mapping rules from XSD to RDF establish that a local element with complex type translates into an object property and a Class. An insight analysis of the META-SHARE schema showed that the straightforward application of such a principle may derive into unnecessary verbose graphs. Thus, following [15], we identified potentially removable nodes before undertaking the actual RDFication process. The criteria applied take into account the tree structure of the nodes, their cardinality and the XPath axes. Thus, embedded complex elements with cardinalityMax=1 are identified as potentially removable, provided they do not contain text nor attributes. This allows for a simplification of the model, for example in the chain `resourceInfo/identificationInfo/resourceName`, the `identificationInfo` property is not needed. Interestingly enough, the removal of the superfluous wrapping elements has also led to a change of philosophy to the schema and a need for restructuring in order to ensure that properties are attached to the most appropriate node, as exemplified and discussed in Section 3.4. Beyond this, we made the following extensions to our mapping strategy:

- renaming of elements when falling into one of the following categories: (a) removal of the `Info` suffix from the wrapping elements of components, as this makes no sense in the new philosophy of the schema: e.g. `validationInfo` becomes validation (as property) and Validation (as class); (b) improvement of names that created confusion, as already noted by the META-SHARE group and/or the ld4lt group; thus, 'resource' was renamed 'languageResource', 'restrictionsOfUse' became 'conditionsOfUse', etc.; (c) as a consequence of the generalization of concepts, e.g. `notAvailableThroughMetashare` with `availableThroughOtherDistributor`; (d) to avoid duplicates created due to the removal of the Info suffix, as described above, e.g. `characterEncoding` becomes `characterEncodingSet` to differentiate from the property `characterEncoding` (previously `characterEncodingInfo`)
- Developement of novel classes based on existing values, e.g., Corpus $\equiv$ $\exists$resourceType.corpus
- Movement of properties to other nodes: obviously, when components are removed, the properties are attached to the higher node; but also for cases where there's a change of concept, as described in 3.4
- Grouping similar elements under novel superclasses, e.g. `annotationType` and `genre` values are structured in classes and subclasses better reflecting the relation between them: the superclass SemanticAnnotation can be used to bring together semantic annotation types, such as semantic roles, named entities, polarity, semantic relations etc.
- Extension of existing classes with new values and including new properties (see section 3.4

3.3 Interface with DCAT and other vocabularies [JPM]

The META-SHARE model can be considered broadly similar to DCAT in that there are classes that are nearly an exact match to ones in DCAT for three out of four cases. DCAT's *dataset* corresponds nearly exactly to the *resource info* tag and similarly, *distributions* are similar to *distribution info* classes and *catalog record* is similar to *metadata info*. The fourth main class, *catalog* covers a level not modelled by META-SHARE. DCAT uses Dublin Core properties for many parts of the metadata, and often these properties are in fact deeply nested into the description. For example, language is found in several places deeply nested under six tags[14]. In META-SHARE this allows different media types in the resource to have different languages ,e.g., the dialogues and the scripts of a video may be in English, but the subtitles can be in French and German (two translations). We still include this fine-grained metadata but also add the property at the resource level to indicate if any part of the resource is in the stated language. Similarly, it also the case that some Dublin Core properties are not directly specified in the META-SHARE model, but can be inferred from related properties, e.g., Dublin Core's 'contributor' follows from people indicated as 'annotators', 'evaluators', 'recorders' or 'validators'. Similarly, several DCAT specific-properties, such as 'download URL', are nearly exactly equivalent to those in Metashare but occur in places that do not fit the domain and range of the properties. In this particular case, it was a simple fix to move the property to the enclosing `DistributionInfo` class. Inevitably, several properties from DCAT did not have equivalences in META-SHARE, notably 'keyword'.

3.4 Licensing module [VRD, PL]

One of the most important achievements of META-SHARE has been the formulation of a clear, consise and easy-to-use licensing model to specify the rights information of the LRs.

In order to limit fuzziness in the terms and conditions of use of LRs, META-SHARE recommends the use of standard licenses (preferably open ones); while proprietary or closed licenses and texts with terms of use are to be avoided. Moreover, the metadata schema includes a module on licensing, which forces LR providers to document the conditions of use of their resources in a standardised format. Elements in the module encode rights holders, the most frequently used conditions and terms (e.g. attribution, no derivatives etc.), formats and location of the distribution files, pricing details etc., while the licence itself is obligatorily selected from a list of predefined values representing the recommended licences.

In the conversion of META-SHARE from XSD to OWL/RDF, we decided to replace the components with classes that can be used to better represent the licensing ecosystem of LRs, and to re-structure the elements in order to make clearer the properties associated with them. As a result, we recognize the following three entities/classes, each associated with different properties as appropriate:

- `LanguageResource`, which is the intellectual property *work*, can be attributed with the `iprHolder`, `distributionRightsHolder`;
- `Distribution`, taken from the DCAT vocabulary where it "*represents an accessible form of a dataset as for example a downloadable file, an RSS feed or a web service that provides the data*"; this is the entity to which properties for describing licencing, forms and other details of distribution must be attached;

[14] `resourceInfo` > `resourceComponentType` > `corpus`* > `corpusMediaType` > `corpusVideoInfo` > `languageInfo`

– `License`, with the specific information that can help us generalize over terms and conditions and enriched with concepts from the ODRL ontology.

Terms and conditions of use can be declared by using URIs pointing to the legal text (or a human readable version) of well-known licenses or to a URL with the specific terms of the LR provider. However, this practice does not favour automated processing and the rights information thus referred would not be queriable. In order to overcome this, a fine-grained representation of licenses, where the specific rights and conditions are given in RDF, was decided. Some languages already exist for this purpose, and among them, ODRL 2.1 was chosen and extended. ODRL (Open Digital Rights Language) is a policy and rights expression language specified by the W3C ODRL Community Group[15] which defines a model for representing permissions, prohibitions and duties, as well as a core vocabulary. The most common licenses (for software, data or general works) have been already expressed in ODRL in the RDF License dataset[13] and can be pointed to when an LR is licensed with any of these.

Extensions to the vocabulary were needed to represent some of the specificities of the LRs domain. The specification and led to changes, some of them structural, with respect to the previous versions. These changes included the selection of classes and properties from other existing vocabularies (specifically from ODRL, Dublin Core, Creative Commons REL[16] and SKOS) as well as the definition of new ones. In addition, the recommended standard licenses have been represented in ODRL and published[17] of the RDF resources to describe licenses was based on a list of requirements[18]. Further, as a support for the representation of non-standard licences (that has to be done by Semantic Web laymen), the new concept of *license templates* has been proposed. A license template is an RDF document with common terms and conditions (e.g. attribution) mapped to ODRL actions (duty to attribute) which are ready to be complemented by other information that changes more frequently. In this way, some of the variable elements are detached and more easy editable.

3.5 Harmonizing other resources with META-SHARE [JPM]

The LingHub portal indexes metadata from a wide-range of sources. While a basic level of interoperability can be established by used standard vocabularies such as DCAT and Dublin Core, this can only be done by sacrificing completeness and ignoring all metadata particular to language resources. For this reason, we rely the META-SHARE model to represent and harmonize the metadata relating specifically to the domain of linguistics and language resources. As a proof-of-concept, we show how the META-SHARE ontology supports the harmonization of CLARIN data. The CLARIN repository describes its resources using a small common set of metadata and a larger description defined by the

[15] https://www.w3.org/community/odrl/
[16] http://creativecommons.org/ns
[17] http://purl.org/NET/ms-rights
[18] https://www.w3.org/community/ld4lt/wiki/Metasharevocabularyforlicenses

Component Metadata Infrastructure [4, CMDI]. These metadata schemes are extremely diverse as shown in table 1. We will focus on the top five of these types, where we have also developed mappings using the LIXR model. Two of these schemes are only Dublin Core properties and so do not have specific language resource metadata. The most frequent 'Song' tag focusses on a database of musical recordings, and many of these properties (e.g., 'number of stanzas') did not correspond to any properties, however the META-SHARE Ontology could be used to describe the language and technical format information (i.e., 'audio encoding'). The `Session` tag is in fact the IMDI metadata [3] and as such corresponds loosely with META-SHARE but highlighted areas where the META-SHARE ontology does not provide sufficient properties, for example in describing the participants in a media recording. The MODS metadata scheme [**?**] was similar in that the META-SHARE ontology provided some properties but was often insufficient in the details that were recorded. This highlights the advantage of taking an open world, ontological approach as opposed to a fixed schema, in that we can easily introduce new properties while still reusing the META-SHARE properties where they were available. I doubt I will manage it but I will try to include the number of MS props used - JPM

Component Root Tag	Institutes	Frequency
Song	1 (MI)	155,403
Session	1 (MPI)	128,673
OLAC-DcmiTerms	39	95,370
mods	1 (Utrecht)	64,632
DcmiTerms	2 (BeG,HI)	46,160
SongScan	1 (MI)	28,448
media-session-profile	1 (Munich)	22,405
SourceScan	1 (MI)	21,256
Source	1 (MI)	16,519
teiHeader	2 (BBAW, Copenhagen)	15,998

Table 1. The top 10 most frequent component types in CLARIN and the institutes that use them. Abbreviations: MI=Meertens Institute (KNAW), MPI=Max Planck Insitute (Nijmegen), BeG=Netherlands Institute for Sound and Vision, HI=Huygens Institute (KNAW), BBAW=Berlin-Brandenburg Academy of Sciences

4 Applications

4.1 IULA LOD Catalogue [MV]

The IULA-UPF CLARIN Competence Centre[19] aims to promote and support the use of technology and text analysis tools in the Humanities and Social Sci-

[19] http://www.clarin-es-lab.org/index-en.html

ences research. The centre includes a Catalogue[20] with information on language resources and technology. The Catalogue is based on the initial LOD version of the META-SHARE model as described in [15] and the original data come from the UPF META-SHARE node[21]. The source XML records were converted into RDF and augmented with service descriptions (not included in the UPF META-SHARE node) and relevant documentation (appropriate articles, documentation, sample data and results, illustrative experiments, examples from outstanding projects, illustrative use cases, etc) to encourage potential users to embrace digital tools. Finally, the data was enriched with internal and external links. The eventual linked data allowed maximizing the information contained in the original repository and developing data mashup techniques that get relevant data from the DBpedia and the DBLP[22]. The Catalogue demonstrates the benefits of the LOD framework and how this can be easily used as the basis for a web browser application that maximizes information and helps users to navigate throughout the dataset in a comprehensive way.

4.2 LingHub [JPM]

5 Conclusion [PC, JPM]

This work represents only a first starting point for the harmonization of language resources by providing a standard ontology that can be used in the description of metadata of linguistic resources. The LingHub portal we have presented here is proof-of-concept for the level of harmonization that the use of a common ontology provides, as metadata originating from different repositories can be uniformly queried in LingHub in an integrated fashion. We adhere to an open architecture in which not only LingHub but other discovery services aggregate and index data could potentially be developed. The work described here is only a first step to harmonization in that there are still a number of challenges ahead of us to be addressed:

- **Data availability:** The next step would be to make sure that not only metadata, but the actual data is available on the Web in open web standards such as RDF so that data can be automatically crawled and analyzed.
- **Data integration and querying:** Linguistic data published on the Web should ideally follow the same format (e.g. RDF) so that it can be easily integrated and data can be queried across datasets. This presupposes the agreement on best practices for data publication and formats. The Natural Language Processing Interchange Format (NIF)[9] is an obvious candidate for that.
- **Service harmonization and discovery:** Harmonization should be extended to the description of NLP services so that NLP services can be dissevered across providers and repositories. The mechanisms for description of the functionality of NLP services should be extremely light-weight.

[20] http://lod.iula.upf.edu/
[21] http://metashare.upf.edu
[22] http://dblp.uni-trier.de/db/index.html

– **Service composition and execution on the cloud:** Input and output formats for services should be standardized and homogenized so that services can be easily composed to realize more complex workflows, without relying on too much parametrization. Workflows of services should be easily executable *'on the cloud'*. In order to scale, services should support parallelization and streaming and support non-centralized processing. Service execution and composition should not require special libraries, grids or other proprietary infrastructures or protocols, but rely only on open web standards and protocols such as the hypertext transfer protocol (HTTP) and content negotiation, ideally being RESTful to keep APIs simple and stateless.

Acknowledgments. We are very grateful to the members of the W3C Linked Data for Language Technologies (LD4LT) for all the useful feedback received and for allowing this initiative to be developed as an activity of the group. This work is supported by the FP7 European project LIDER (610782), by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R) and the Greek CLARIN Attiki project (MIS 441451).

References

1. Bird, S., Simons, G.: The OLAC metadata set and controlled vocabularies. In: Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources-Volume 15. pp. 7–18. Association for Computational Linguistics (2001)
2. Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C.: A data category registry-and component-based metadata framework. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. pp. 43–47 (2010)
3. Broeder, D., Offenga, F., Willems, D., Wittenburg, P.: The IMDI metadata set, its tools and accessible linguistic databases. In: Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia. pp. 11–13 (2001)
4. Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., Trippel, T.: CMDI: a component metadata infrastructure. In: Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme. p. 1 (2012)
5. Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., Soria, C.: The LRE Map. Harmonising community descriptions of resources. In: Proceedings of the Eighth Conference on International Language Resources and Evaluation. pp. 1084–1089 (2012)
6. Chiarcos, C.: Ontologies of linguistic annotation: Survey and perspectives. In: LREC. pp. 303–310 (2012)
7. Farrar, S., Lewis, W., Langendoen, T.: A common ontology for linguistic concepts. In: Proceedings of the Knowledge Technologies Conference. pp. 10–13 (2002)
8. Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., Mapelli, V.: The META-SHARE metadata schema for the description of language resources. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 1090–1097 (2012)

9. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: The Semantic Web–ISWC 2013, pp. 98–113. Springer (2013)
10. Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: ISOcat: Corralling data categories in the wild. In: LREC (2008)
11. Motik, B., Patel-Schneider, P.F., Parsia, B., Bock, C., Fokoue, A., Haase, P., Hoekstra, R., Horrocks, I., Ruttenberg, A., Sattler, U., Smith, M.: OWL 2 web ontology language structural specification and functional-style syntax. W3C recommendation, The World Wide Web Consortium (2012)
12. Piperidis, S.: The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In: Proceedings of the Eighth Conference on International Language Resources and Evaluation. pp. 36–42 (2012)
13. Rodriguez-Doncel, V., Villata, S., Gomez-Perez, A.: A dataset of rdf licenses. In: Proceedings of the 27th Int. Conf. on Legal Knowledge and Information System (JURIX). pp. 187–189 (2014)
14. Ďurčo, M., Windhouwer, M.: From CLARIN component metadata to linked open data. In: Proceedings of the 3rd Workshop on Linked Data in Linguistics. pp. 13–17 (2014)
15. Villegas, M., Melero, M., Bel, N.: Metadata as linked open data: mapping disparate xml metadata registries into one rdf/owl registry. In: Chair), N.C.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). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)