

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

John P. McCrae¹, Penny Labropoulou³, Jorge Gracia², Marta Villegas⁴, Víctor Rodríguez-Doncel², and Philipp Cimiano¹

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. META-SHARE is a repository where significant effort has been made into providing carefully curated metadata about language resources (LRs). However, in the face of the flood of data that is used in computational linguistics, a manual approach cannot suffice. We present the development of the META-SHARE ontology, which transforms the rigid metadata schema used by META-SHARE so far into an open world ontology that can better handle the diversity of metadata found in legacy and crowd-sourced resources. We show how this model can interface with other more general purpose vocabularies for online datasets and licensing, and apply this model to the CLARIN VLO, a large source of legacy metadata about LRs. Furthermore, we demonstrate the usefulness of this approach in two public metadata portals for information about language resources.

Keywords: language resources and evaluation, metadata, ontologies, harmonization

1 Introduction

The study of language and the development of natural language processing (NLP) applications requires 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 notable of such initiatives is META-SHARE⁵ [16], an open, integrated, secure and interoperable exchange infrastructure where LRs are documented, uploaded, stored, catalogued, announced, downloaded, exchanged and discussed, aiming to support reuse of LRs.

⁵ <http://www.meta-share.eu>

Towards this end, META-SHARE has developed a rich metadata schema that allows aspects of LRs accounting for their whole lifecycle from their production to their usage to be described. The schema has been implemented as an XML Schema Definition (XSD) ⁶ and descriptions of specific LRs are available as XML documents.

Yet, META-SHARE is not the only metadata repository for LRs; other repositories include the CLARIN Virtual Language Observatory (VLO)⁷ [2] as well as the LRE-Map⁸ [5]. The metadata schemes of these different repositories vary with respect to their coverage and the set of specific metadata captured. 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. In this paper we contribute to the interoperability of these repositories by developing an ontology in the Web Ontology Language (OWL) [15] that allows us to represent the metadata schemes of these repositories under an extensible, open-world model.⁹ The proposed ontology is based on the ontology developed by Villegas et al. [19] for the University Pompeu Fabra's (UPF) META-SHARE node, covering part of the original schema, but extended it to the complete schema and to cover all relevant LRs by incorporating the consensus reached in the context of the W3C Linked Data for Language Technologies (LD4LT) Community Group¹⁰. We show how this model interacts with the DCAT [14] vocabulary as well as the most prominent models in the CLARIN VLO data. Further, we describe the application of the model in two portals, firstly the IULA LOD catalogue and secondly *Linghub*¹¹. Our approach has several advantages. Firstly, the use of Semantic Web techniques support standardized means of representing, linking, and accessing the data, while the resulting data is lighter, better suited for exploitation and eases further extensions and links with external resources. In addition, we assume 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. We evaluate this hypothesis by reference to CLARIN.

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 and its application in Section 4. 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.

⁶ <https://github.com/metashare/META-SHARE/tree/master/misc/schema/v3.0>

⁷ <http://catalog.clarin.eu/vlo/?1>

⁸ <http://www.resourcebook.eu/searchll.php>

⁹ <http://purl.org/net/def/metashare>

¹⁰ <https://www.w3.org/community/ld4lt>

¹¹ <http://linghub.org/>

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 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]. This lack of consensus resides also in the description of LRs, even for non-linguistic concepts. In fact, there are as many metadata schemas for their descriptions as catalogs and repositories for their presentation (e.g. those used by ELRA and the LDC) and communities describing them (e.g. TEI [12] or CES [11]). The most widely used schema for the exchanges of LRs is the one suggested by the Open Language Archives Community [1, OLAC], which builds on the Dublin Core metadata and has been criticized as too reductionistic. Differences between the schemas lie in the range of features included (e.g. size, technical details, classification etc.), the labels used for the features (e.g. rights, license, policy, terms of use etc. to denote the terms of use for an LR) and the datatypes for the features (e.g. values taken from a closed vocabulary or input as free text).

An important effort to harmonize metadata has been the ISO Data Category Registry (ISOcat DCR) [13], intended as a registry where metadata providers could register their concepts (Data Categories) and link them to those of other providers. A subset thereof were selected by metadata experts as the core elements for the description of LRs. Extending this ISOcat principle, the Component Metadata Infrastructure [4] proposed by CLARIN attempts to bring together “components” consisting of semantically close elements to be shared among different communities when producing “profiles” for specific LR types. 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 [18]. A different approach was taken for the design of the META-SHARE schema [9], which was based on a comparative study of metadata schemas, analysis of user needs and discussions with metadata providers and experts in order to arrive at a common schema.

3 The META-SHARE OWL Ontology

3.1 Original MS XSD schema

The META-SHARE schema [9] has been designed not only as an aid for LR search and retrieval, but also as a means to foster their production, use and reuse 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. The central entity of the META-SHARE schema is the LR *per se*, which

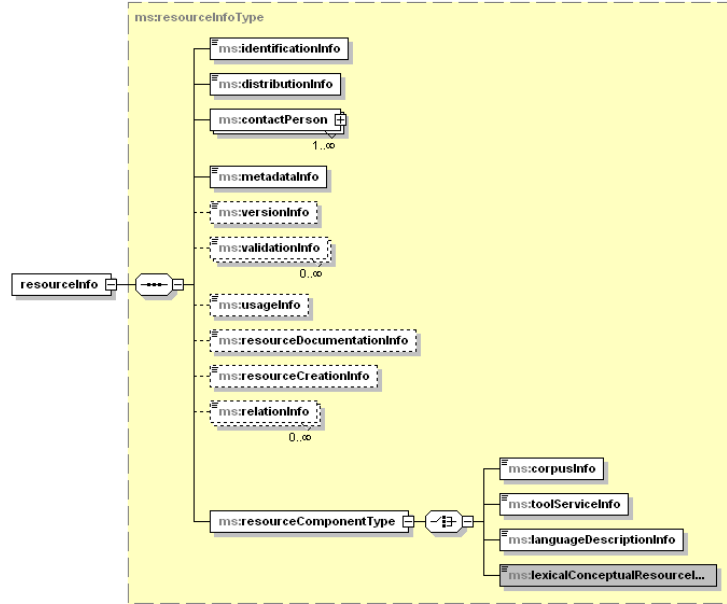


Fig. 1. The core of the META-SHARE model

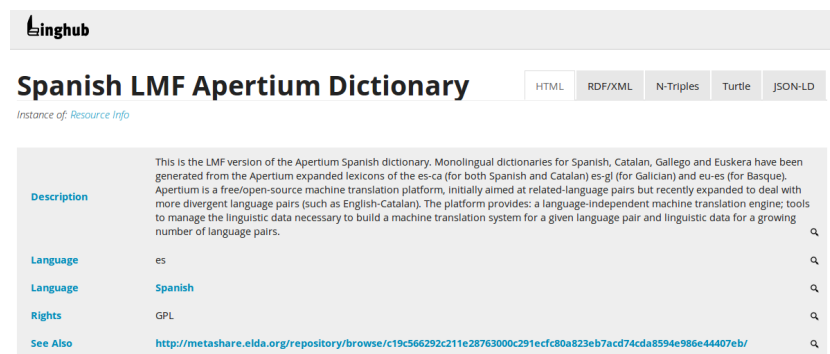
encompasses both **data sets** (e.g., textual, audio and multimodal/multimedia corpora, lexical data, ontologies, terminologies, computational grammars, language models) and **technologies** (e.g., tools, services) used for their processing. In addition to the central entity, other entities are also 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). 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), which subsumes

- administrative components relevant to all LRs, e.g. **identificationInfo** (name, description and identifiers), **distributionInfo** (licensing and intellectual property rights information), **usageInfo** (information about the intended and actual use of the LR).
- components specific to the resource type (corpus, lexical/conceptual resource, language model, tool/service) and media type (text, audio, video, image), which support the encoding of information relevant to resource/media

combinations, e.g. text or audio parts of corpora, lexical/conceptual resources etc., such as language, formats, classification.

The META-SHARE schema recognises obligatory elements (minimal version) and recommended and optional elements (maximal version). 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



The screenshot shows the Linghub interface for the 'Spanish LMF Apertium Dictionary'. At the top, there's a Linghub logo and a title bar. Below the title, there are tabs for different formats: HTML (selected), RDF/XML, N-Triples, Turtle, and JSON-LD. A sub-header 'Instance of: Resource Info' is visible. The main content area is a table with the following rows:

Description	This is the LMF version of the Apertium Spanish dictionary. Monolingual dictionaries for Spanish, Catalan, Gallego and Euskera have been generated from the Apertium expanded lexicons of the es-ca (for both Spanish and Catalan) es-gl (for Galician) and eu-es (for Basque). Apertium is a free/open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan). The platform provides: a language-independent machine translation engine; tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.	
Language	es	
Language	Spanish	
Rights	GPL	
See Also	http://metashare.eida.org/repository/browse/c19c566292c211e28763000c291ecf80a823eb7acd74cda8594e986e44407eb/	

Fig. 2. META-SHARE data as displayed within the Linghub interface

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. We observed that the straightforward application of such a principle may derive unnecessarily verbose graphs. Thus, following Villegas et al. [19], we identified potentially removable nodes before undertaking the actual RDFication process. Embedded complex elements with cardinality of exactly one are identified as potentially removable, provided they contain neither text nor attributes. This allows for a simplification of the model, for example in the chain `resourceInfo o identificationInfo o resourceName`, the `identificationInfo` property is not needed. Interestingly enough, the removal of the superfluous wrapping elements has also led to a change of philosophy in 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 extensions to our mapping strategy in order to improve the ontology, such as the following:

- Removal of the **Info** suffix from the names of wrapping elements of components.
- Improvement of names that created confusion, as already noted by the META-SHARE group and/or the LD4LT group; thus, **resourceInfo** was renamed **LanguageResource**, **restrictionsOfUse** became **conditionsOfUse**.
- Generalization of concepts, e.g. **notAvailableThroughMetashare** with **availableThroughOtherDistributor**;
- Development of novel classes based on existing values, for example:
 $\text{Corpus} \equiv \exists \text{resourceType.corpus}$
- Grouping similar elements under novel superclasses, e.g. **annotationType** and **genre** values are structured in classes and subclasses better reflecting the relation between them. Indicatively, the superclass **SemanticAnnotation** can be used to bring together semantic annotation types, such as semantic roles, named entities, polarity, and semantic relations.
- Extension of existing classes with new values and new properties (e.g. **licenseCategory** for licences).

The actual mapping was achieved by means of a custom domain-specific language inspired by XML Stylesheet Transforms.

3.3 Interface with DCAT and other vocabularies

The META-SHARE model can be considered broadly similar to DCAT in that there are classes that are nearly an exact match to the ones in DCAT for three out of four classes. DCAT's **dataset** corresponds nearly exactly to the **resourceInfo** tag and similarly, **distributions** are similar to **distributionInfo** classes and **catalogRecord** is similar to **metadataInfo**. Thus, we introduced *equivalent class* relations between these elements. 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 found deeply nested in the META-SHARE description. For example, language is found in several places deeply nested under six tags¹². 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, whereas the subtitles can be in French and German. 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 is 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 (by means of a property chain) from people indicated as 'annotators', 'evaluators', 'recorders' or 'validators'. Furthermore, 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

¹² e.g., $\text{resourceInfo} \circ \text{resourceComponentType} \circ \text{corpus} \circ \text{corpusMediaType} \circ \text{corpusVideoInfo} \circ \text{languageInfo} \rightarrow \text{dc:language}$

```

<> a dcat:Dataset ;
    dc:description "Cette base de donnees..."@fr
    dc:language "tur" ;
    dc:source "META-SHARE" ;
    dc:title "Turkish Continuous and Isolated Word Speech Database"@en ;
    dct:type ms:corpus ;
    ms:corpusInfo <#corpusInfo> ;
    ms:distributionInfo <#distributionInfo> ;
    rdfs:seeAlso <http://metashare.elda.org/reposit...ac770/> .

<#corpusInfo> a ms:CorpusInfo , ms:CorpusAudioInfo ;
    dc:language "tur" ;
    ms:languageName "Turkish" ;
    ms:mediaType ms:audio ;
    ms:linguality ms:monolingual .

<#distributionInfo> a ms:DistributionInfo , dcat:Distribution ;
    dct:license <#licenceInfo> , <#licenceInfo2> ;
    ms:availability ms:restrictedUse ;
    dcat:accessURL <http://catalog.elra...> .

```

Fig. 3. An abridged example of a metadata entry represented with common metadata properties from DCAT and Dublin Core and novel properties from the MetaShare ontology.

fix to move the property to the enclosing `DistributionInfo` class. Inevitably, several properties from DCAT did not have equivalences in META-SHARE, notably ‘keyword’.

In addition to DCAT, we used also other vocabularies to establish equivalences to parts of the model. In particular, we mapped to the Friend of a Friend (FOAF) ontology to describe people and organizations and the Semantic Web for Research Communities (SWRC) ontology to describe scientific publications.

3.4 Licensing module

A specific area where we made a significant effort to improve the modelling was in the licensing information in order to allow the formulation of a clear and concise rights information of the LRs. Some languages already exist for this purpose, and among them, ODRL 2.1 (Open Digital Rights Language) was chosen and extended, which is a policy and rights expression language specified by the W3C ODRL Community Group¹³ which defines a model for representing permissions, prohibitions and duties. The most common licenses (for software, data or general works) have been already expressed in ODRL in the RDF License dataset [17] and can be pointed to when an LR is licensed with any of these. Extensions to the ODRL vocabulary have been made to represent some of the

¹³ <https://www.w3.org/community/odrl/>

specificities of the LR domain. The specification also suggested changes, some of them structural, to the previous META-SHARE modelling, and to this extent we combined the existing META-SHARE licensing vocabulary with ODRL. In addition, we extended the model by adding some new properties and individuals based on requirements from the LD4LT community group¹⁴.

META-SHARE limits fuzziness in the terms and conditions of use of LRs by providing a range of recommended standard licenses organized on the following axes: open licenses are the preferred option (Creative Commons licenses for data resources and Free Open Source Software for tools and services), followed by two sets of model (standard) licenses built in response to LR providers' requests (META-SHARE Commons and NoRedistribution licenses); custom and proprietary licenses are also allowed for legacy resources that cannot be licensed otherwise. Further, the generic *conditions-of-use* values of the META-SHARE schema have been exploited for creating RDF codes for non-standard licenses: common terms and conditions (e.g. attribution) are mapped to ODRL actions (e.g. the duty to attribute), and included in an RDF document. Extension with more variable aspects of licenses can be added on top. In this way, some of the variable elements are detached and more easily editable.

This module has been published both as independent module¹⁵ and as part of the META-SHARE ontology.

3.5 Harmonizing other resources with META-SHARE

While a basic level of interoperability can be established by using 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 use 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 data from the CLARIN VLO. The CLARIN repository describes its resources using a small common set of metadata and a larger description defined by the 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 for which we have created corresponding mappings. Two of these schemes are only Dublin Core properties and so do not have specific language resource metadata. The most frequent tag 'Song' tag is used to describe records of a database consisting of musical recordings. While many of the properties used by this tag (e.g., 'number of stanzas') have no correspondence in Dublin Core, they can be described with respect to the META-SHARE Ontology through properties to describe the technical format information (i.e., 'audio encoding'). The **Session** tag is used to provide IMDI metadata [3] and has a very loose correspondence to META-SHARE: For instance, there are no corresponding properties to describe the participants of a media recording. This highlights

¹⁴ https://www.w3.org/community/ld4lt/wiki/Metashare_vocabulary_for_licenses

¹⁵ <http://purl.org/NET/ms-rights>

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 are appropriate. The MODS metadata scheme [8] is in fact a general domain metadata framework. We found that 28 entities from META-SHARE corresponded to elements used in the MI ‘Sing’ metadata, and 37 to the IMDI metadata, although there was only minor overlap with the MODS scheme (in particular 4 entities used to describe language) as this scheme is not specific to language resources.

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

The IULA-UPF CLARIN Competence Centre¹⁶ aims to promote and support the use of technology and text analysis tools in the Humanities and Social Sciences research. The centre includes a Catalogue¹⁷ with information on language resources and technology. The Catalogue is based on the initial linked open data (LOD) version of the META-SHARE model as described in [19] and the original data originate from the UPF META-SHARE node¹⁸. 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

¹⁶ <http://www.clarin-es-lab.org/index-en.html>

¹⁷ <http://lod.iula.upf.edu/>

¹⁸ <http://metashare.upf.edu>

users to embrace digital tools. Finally, the data was enriched with internal and external links. The resulting linked data maximised the information contained in the original repository and enabled data mashup techniques that get relevant data from the DBpedia and the DBLP¹⁹. 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 datasets in a comprehensive way.

4.2 Linghub

Linghub is a portal designed to allow common querying of metadata from multiple highly heterogeneous repositories. Currently, this repository draws not only from META-SHARE, but also from the LRE-Map [5], the CLARIN VLO [2] and DataHub. The repository currently bases itself mostly on the DCAT and Dublin Core vocabularies, however these models do not capture any specific linguistic information. For this reason, the ontology described in this paper is currently being integrated into the system to allow users to use META-SHARE as the basic vocabulary for querying linguistic information about language resources, and the mappings previously described have already been applied to data from LRE-Map and the CLARIN VLO. Linghub supports browsing and querying by several means, including faceted browsing, full-text search, SPARQL querying and related item search. As such, we believe that the portal, while not a direct collector of metadata, will enable users to find more language resources and do so more easily. The Linghub portal is thus a 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.

5 Conclusion

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 and there are still a number of challenges ahead of us to be addressed. Firstly, 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. Secondly, it should be required that 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 and the Natural Language Processing Interchange Format (NIF)[10] is an obvious candidate for that. Thirdly, harmonization should be extended to the description of NLP services so that

¹⁹ <http://dblp.uni-trier.de/db/index.html>

NLP services can be distributed across providers and repositories. The mechanisms for description of the functionality of NLP services should be extremely light-weight. Finally, 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, streaming and non-centralized processing. We believe that the development of common vocabularies such as the one presented in paper should enable the emergence of a new paradigm supporting the discovery and exploitation of linguistic data and services across repositories.

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 a Juan de la Cierva grant) 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. 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. pp. 1–4 (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: Proceedings of the Eighth International Conference on Language Resources and Evaluation. 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. Gartner, R.: MODS: Metadata object description schema. JISC Techwatch report TSW pp. 3–6 (2003)
9. 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. pp. 1090–1097 (2012)
10. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: Proceedings of the 12th International Semantic Web Conference, pp. 98–113. Springer (2013)
 11. Ide, N.: Corpus encoding standard: SGML guidelines for encoding linguistic corpora. In: Proceedings of the First International Language Resources and Evaluation Conference. pp. 463–470 (1998)
 12. Ide, N., Véronis, J.: Text encoding initiative: Background and contexts, vol. 29. Springer Science & Business Media (1995)
 13. Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: ISOcat: Corraling data categories in the wild. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2008)
 14. Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (DCAT). W3C recommendation, The World Wide Web Consortium (2014)
 15. 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)
 16. 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)
 17. 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)
 18. 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)
 19. Villegas, M., Melero, M., Bel, N.: Metadata as linked open data: mapping disparate XML metadata registries into one RDF/OWL registry. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. pp. 393–400 (2014)