Reconciling Heterogeneous Descriptions of Language Resources

John P. McCrae, Philipp Cimiano
CIT-EC, Bielefeld University
Bielefeld, Germany
{jmccrae, cimiano}@cit-ec.uni-bielefeld.de

Luca Matteis, Roberto Navigli
University of Rome, La Sapienza
Rome, Italy
{matteis, navigli}@di.uniroma1.it

Victor Rodríguez Doncel, Daniel Vila-Suero
Jorge Gracia
Universidad Politécnica de Madrid
Madrid, Spain
{vrodriguez, dvila, jgracia}@fi.upm.es

Andrejs Abele, Gabriela Vulcu
Paul Buitelaar
Insight Centre, National University of Ireland
Galway, Ireland
{andrejs.abele, gabriela.vulcu,
paul.buitelaar}@insight-centre.org

Abstract

Language resources are a cornerstone of linguistic research and for the development of natural language processing tools, but the discovery of relevant resources remains a challenging task. This is due to the fact that relevant metadata records are spread among different repositories and it is currently impossible to query all these repositories in an integrated fashion, as they use different data models and vocabularies. In this paper we present a first attempt to collect and harmonize the metadata of different repositories, thus making them queriable and browsable in an integrated way. We make use of RDF and linked data technologies for this and provide a first level of harmonization of the vocabularies used in the different resources by mapping them to standard RDF vocabularies including Dublin Core and DCAT. Further, we present an approach that relies on NLP and in particular word sense disambiguation techniques to harmonize resources by mapping values of attributes – such as the type, license or intended use of a resource – into normalized values. Finally, as there are duplicate entries within the same repository as well as across different repositories, we also report results of detection of these duplicates.

1 Introduction

Language resources are the cornerstone of linguistic research as well as of computational linguistics. Within NLP, for instance, most tools developed require a corpus to be trained (e.g. language models, statistical taggers, statistical parsers, and statistical machine translation systems) or they require lexico-semantic resources as background knowledge to perform some task (e.g. word sense disambiguation). As the number of language resources available keeps growing, the task of discovering and finding resources that are pertinent to a particular task becomes increasingly difficult. While there are a number of repositories that collect and index metadata of language resources, such as META-SHARE (Federmann et al., 2012), CLARIN (Broeder et al., 2010), LRE-Map (Calzolari et al., 2012), Datahub.io[1] and OLAC (Simons and Bird, 2003), they do not provide a complete solution to the discovery problem for two reasons. First, integrated search over all these different repositories is not possible, as they use different data models, different vocabularies and expose different interfaces and APIs. Second, these repositories must strike a balance between quality and coverage, either opting for coverage at the expense of quality of metadata, or *vice versa*.

When collecting metadata from multiple resources, we understand that there are two principal challenges: property harmonization and duplication detection. Harmonization is the challenge of verifying that there is not only structural and syntactic interoperability between the resources in that they use the same property, for example Dublin Core's language property, but also that they use the same value. For example, the following values of the language property are likely to be equivalent: "French", "Modern French", "français", "fr", "fra" and "fre". It is difficult to write queries on a dataset if every property has many equivalent values and thus it is essential to use a single representation. Secondly, we wish to

[1] http://datahub.io/

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 39–48,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

detect duplicates that occur either due to the original representation or from multiple sources. It is clear that if a large number of records in fact describe the same resource then queries for that resource will return too many resources that may lead to errors (or annoyance) for users. For example, the "Universal Declaration of Human Rights" is available in 444 languages[2] and listing each translation as a single resource (as the CLARIN VLO does) does not correctly capture the nature of the resource. Furthermore, these resources may not match some queries, such as for example 'resources in more than one language', and as such it is preferable to merge these individual records into a single complex record.

As the main contribution of this paper, we present the methods used to harmonize data across repositories. Due to the different kinds of values and target taxonomies chosen for each property, these methods vary but all are based on state-of-the-art NLP techniques, including word sense disambiguation, and make major improvements to the data quality of our metadata records. Second, we show indeed that duplicate metadata records are pervasive and that they occur both within and across repositories. We then present a simple yet effective approach to detect duplicates within and across repositories.

The paper is structured as follows: we give an overview of work related to harmonization of data as well as an overview of existing metadata repositories for linguistic data in Section 2. We describe our metadata collection and schema matching strategy in Section 3. We describe our techniques for metadata harmonization in Section 4. We describe our methods for duplication detection in Section 5. The performance of the different techniques is reported in each of these sections. We discuss our methodology and approach from a wider point of view in Section 6.

2 Related Work

Interoperability of metadata is an important problem in many domains and harmonizing schemas from different sources has been recognized as a major challenge (Nilsson, 2010; Khoo and Hall, 2010; Nogueras-Iso et al., 2004). There are different approach to data integration. One approach consists on mapping data to one *monolithic* ontology that needs to be general enough to accommodate all the data categories from different sources. While this is appealing as it supports integrated querying of data, a single ontology cannot predict all aspects of metadata records, that all users may wish to record. In contrast, the linked data approach relies on multiple, standardized smaller and reusable vocabularies, each representing a subset of the data. In this line, some experts have recommended (Brooks and McCalla, 2006):

> "A larger set of ontologies sufficient for particular purposes should be used instead of a single highly constrained taxonomy of values."

In the context of linguistic data, different approaches have been pursued to collect metadata of resources. Large consortium-led projects and initiatives such as the CLARIN projects and META-NET have attempted to create metadata standards for representing linguistic data. Interoperability of the data stemming from these two repositories is however severely limited due to incompatibilities in their data models. META-SHARE favors a qualitative approach in which a relatively complex XML schema is provided to describe metadata of resources (Gavrilidou et al., 2012). At the same time, considerable effort has been devoted to ensuring data quality (Piperidis, 2012). In contrast, CLARIN does not provide a single schema, but a set of 'profiles' that are described in a schema language called the *CMDI Component Specification Language* (Broeder et al., 2012). Each institute describing resources using CMDI can instantiate the vocabulary to suit their particular needs. Similarly, an attempt has been made to catalogue language resources by assigning them a single unique identifier (Choukri et al., 2012).

Other more decentralized approaches are found in initiatives such as the LRE-Map (Calzolari et al., 2012) which provides a repository for researchers who want to submit the resources accompanying papers submitted to conferences. Most fields in LRE-Map consist of a text field with some prespecified options to select and a thorough analysis of the results has been conducted (Mariani et al., 2014).

Similarly, the *Open Linguistics Working Group* (Chiarcos et al., 2012) has been collecting language resources published as linked data in a

[2] http://www.ohchr.org/en/udhr/pages/introduction.aspx

Source	Records	RDF Triples	Triples per Record
META-SHARE	2,442	464,572	190.2
CLARIN VLO	144,570	3,381,736	23.4
Datahub.io	218	10,739	49.3
LRE-Map (LREC 2014)	682	10,650	15.6
LRE-Map (Non-open)	5,030	68,926	13.7
OLAC	217,765	2,613,183	12.0
ELRA Catalogue	1,066	22,580	21.2
LDC Catalogue	714	n/a	n/a

Table 1: The sizes of the resources in terms of number of metadata records and total data size

crowd-sourced repository at Datahub.io, in order to monitor the *Linguistic Linked Data cloud* and produce a diagram showing the status of these resources.

This clearly shows that the field is very fragmented, with different players using different approaches and most importantly different meta- and data models, thus impeding the discovery and integration of linguistic data.

3 Metadata collection and Schema Matching

In this section we describe the different methods applied to collect metadata from the different repositories:

- **META-SHARE:** For META-SHARE, a dump of the data was provided by the ILSP managing node of the META-NET project in XML format. We developed a custom script to convert this into the RDF data model, explicitly aligning data elements to the Dublin Core metadata vocabulary and add these as extra RDF triples to the root of the record. Frequently, these properties were deeply nested in the XML file and manual analysis was required to detect which instances truly applied to the entire metadata record.

- **CLARIN:** For CLARIN, we rely on the OAI-PMH (Sompel et al., 2004) framework to harvest data. The harvested OAI-PMH records comprise a header with basic information as well as a download link and a secondary XML description section that is structured according to the particular needs of the data provider. So far, we limit ourselves to collecting only those records that have Dublin Core properties.

- **LRE-Map:** For LRE-Map we used the available RDF/XML data dump[3], which contains submission information from the LREC 2014 conference, as well as data from other conferences, which is not freely available. In the RDF data, we gathered additional information about language resources, including the title of the paper describing the resource.

- **Datahub.io:** The data from Datahub.io was collected by means of the CKAN API[4]. As Datahub.io is a general-purpose catalogue we limited ourselves to extracting only those resources that were of relevance to linguistics. For this, we used an existing list of relevant categories and tags maintained by the Working Group on Open Linguistics (Chiarcos et al., 2012). The data model used by Datahub.io is also based on DCAT, so little adaptation of the data was required.

- **OLAC:** The Open Language Archives Community also relies on OAI-PMH to collect metadata and overlaps significantly with the CLARIN VLO. Unfortunately the data on this site is not openly licensed.

- **ELRA and LDC Catalogues:** These two organizations sell language resources and their catalogues are available online. The metadata records are not themselves openly licensed.

The total size in terms of records and triples (facts) as well as the average number of triples per repository are given in Table 1, where we can see significant differences in size and complexity of the resources. Note for the rest of this paper we will concern ourselves only with the openly licensed resources.

4 Metadata harmonization

As metadata has been obtained from different repositories, there are many incompatibilities between the values used in different resources. While some repositories ensure high-quality metadata in general, we also discovered inconsistencies in the use of values. For instance, while

[3] http://datahub.io/organization/institute-for-computational-linguistics-ilc-cnr

[4] http://datahub.io/api/3/ documented at http://docs.ckan.org/en/latest/api/index.html

META-SHARE recommends the use of ISO 639-3[5] tags for languages, a few data entries use English names for the language instead of the ISO code. We describe our approach to data value normalization below. In this initial harmonization phase we focused on the key questions of whether a resource is available, that is the given URL resolves, and whether the terms and conditions under which the resource can be used are specified. Further, we consider three key aspects that users need to know about resources to help them decide whether the resource matches their needs, namely: the type of the resource (corpus, lexical resource, etc.), intended use of the resource and languages covered. We note that many resources have multiple values for the same property (e.g., language), thus we allow multiple values at the record level, while still permitting more specific annotation deeper in the record.

4.1 Availability

In order to enable applications to (re)use language resources, we should find out if the resources described can still be accessed. For this we focused on the properties which were mapped to DCAT's 'access URL' property in the previous section. These 'access URLs' are intended to refer to HTML pages containing either download links or information on how to retrieve and use the resource. We augment the data with information about which links are valid and about the form of the content returned (e.g. HTML, XML, PDF, RDF/XML, etc.). Therefore, as we deal with heterogeneous sources and repositories, we analyzed access related characteristics and initially focused on answering two questions: *Is the language resource available and accessible on the Web and in what format?*.

To assess the current situation, we crawled and performed an analysis on a set of 119,290 URLs [6]. Our analysis showed that more than 95% of the URLs studied corresponded to accessible URLs (i.e., HTTP Response Code 200 OK), which indicates that in a high number of cases at least some information is provided to potential consumers of the resource.

Furthermore, our assessment showed that more than 66% of the accessible URLs corresponds to HTML pages, around 10% to RDF/XML documents, and other non-text formats sum up to almost 10% of the URLs analyzed (see Table 2). It is important to note that these results only describe what was returned by the service, and do not well reflect the actual format or availability of the data. For example, the high number of resources returning RDF/XML is mostly due to two CLARIN contributing institutes adopting RDF for their metadata.

Format	Resources	Percentage
HTML	67,419	66.2%
RDF/XML	9,940	9.8%
JPEG Image	6,599	6.5%
XML (application)	5,626	5.6%
Plain Text	4,251	4.2%
PDF	3,641	3.6%
XML (text)	3,212	3.2%
Zip Archive	801	0.8%
PNG Image	207	0.2%
gzip Archive	181	0.2%

Table 2: The distribution of the 10 most used formats within the analyzed sample of URLs. Note XML is associated with two MIME types.

4.2 Rights

Language resources are generally protected by copyright laws and they cannot be used against the terms expressed by the rights holders. These terms of use declare the actions that are authorized (e.g. derive, distribute) and the applicable conditions (e.g. attribution, the payment of a fee). They are an essential requirement for the reuse of a resource, but their automatic retrieval and processing is difficult because of the many forms they may adopt: rights information can appear either as a textual notice or as structured metadata, can consist of a mere reference to a well-known license (like an Open Data Commons or Creative Commons license), or it can point to an institution-specific document in a non-English language. These heterogeneous practices prevent the automated processing of licensing information.

Several challenges are posed for the harmonisation of the rights information: first, information is often not legally specified but instead vague statements such as 'freely available' are used; second, description of specific rights and conditions of each license requires complex modelling; and finally, due to the sensitivity of the information,

[5] http://www-01.sil.org/iso639-3/

[6] Due to crawling restrictions, only 60% of the URLs of the dataset were actually crawled

only high precision approaches should be applied.

From the RDF License dataset (Rodriguez-Doncel et al., 2014) we extracted the title, URI and abbreviation of the most commonly used licenses in different forms, and searched for exact matches normalizing for case, punctuation and whitespace. This introduced some errors due to dual-licensing schemes or misleading description were introduced. We manually evaluated all matching licenses and found 95.8% of the recognised strings were correctly matched. With this approach we could identify matching licenses for only 1% of the metadata entries. However, our observations suggest that this is due to the uninformative content for the license attribute. Furthermore, we note that more sophisticated methods have been shown to improve recall, but they do this at the cost of precision (Cabrio et al., 2014).

4.3 Usage

The language resource usage indicates the purpose and application for which the LR was created or which it has since be used. For META-SHARE we rely on the 83 values of the `useNLPSpecific` property and for LRE-Map we have a more limited list of 28 suggested values and many more user-provided free text entries, 3,985 in total (no other source contained this information). We manually mapped the 28 predefined values in LRE-Map to one of the 83 values predefined in META-SHARE. For the user-provided intended usage values, we developed a matching algorithm that identifies the corresponding META-SHARE intended use values. First we tokenized the expressions, then we stemmed the tokens using the Snowball stemmer (Porter, 2001), and we performed a string inclusion match, i.e. checking whether META-SHARE usages are included in the free text entries. For some entries we retrieved several matches (e.g. 'Document Classification, Text categorisation' matched both 'document classification' and 'text categorisation'), assuming that in the case of multiple matches the union of the intended usages was meant. With this algorithm we identified 66 matches on a random sample of 100 user-provided entries and they were all correct matches. From the remaining 34 unmatched entries, 16 were empty fields or non specific e.g. 'not applicable', 'various uses'. Other 16 entries were too general to be mapped to an intended use defined in the META-SHARE vocabulary e.g. 'testing', 'acquisition'. We had one false negative 'taggin pos'[sic] and one usage that is not yet in META-SHARE 'semantic system evaluation'. On this basis we had 98-99% accuracy on the results. Following the aforementioned algorithm we identified 65% matches on the entire set of user-entries. We further investigated the remaining 35% non-matches and we identified further intended use values that are not yet in META-SHARE vocabulary, e.g. 'entity linking', 'corpus creation', which we will suggest as extensions of the META-SHARE vocabulary.

4.4 Language

To clean the names of languages contained in metadata records, we aligned to the ISO 639-3 standard. First we extracted all the language labels from our records and obtained a total of 833 distinct language labels. Next we leveraged two resources to map these noisy language labels to standard ISO codes: (i) the official SIL database[7], which contains all the standard ISO codes and their *English* names, and (ii) BabelNet[8] (Navigli and Ponzetto, 2012), a large multilingual lexico-semantic resource containing, among others, translations and synonyms of various language names along with their ISO codes.

To perform the mapping in an automatic manner, we compared each of the 833 noisy language labels against the language labels contained in SIL and BabelNet using two string similarity algorithms: the *Dice coefficient* string similarity algorithm and the *Levenshtein* distance string metric.

Table 3 reports an excerpt of the results showcasing in the first row a match for all cases, in the second a match for BabelNet but not for SIL, and in the third a mismatch for all. Furthermore, the final row reports a mismatch from Levenshtein, where 'Turkish, Crimean' is matched instead.

In order to measure the accuracy of each approach we tested the mapping algorithms against a manually annotated dataset containing 100 language labels and ISO codes. In Table 4, we present the accuracy of our methods based on the number of labels correctly identified ("label accuracy") and the accuracy weighted for the number of metadata records with that label ("instance accuracy"). The best results are obtained using BabelNet as the source of language labels. Babel-

[7] `http://www-01.sil.org/iso639-3/download.asp`

[8] `http://babelnet.org/`

Input	Expected output	BabelNet output		SIL output	
		dice	*leven*	*dice*	*leven*
Kurdish	kur	kur	kur	kur	kur
rank – distance		1	0	1	0
label		Kurdish	Kurdish	kurdish	Kurdish
Bokmål	nob	nob	nob	bok*	bdt*
rank – distance		1	0	0.57	3
label		bokmål	Bokmål	bok	Bokoto
Ñahñú (Otomí)	oto	omq*	otm*	ttf*	las*
rank – distance		0.38	8	0.35	7
label		Otomí Mangue	Eastern Otomí	tuotomb	lama (togo)
Turkish (Türkçe)	tur	tur	tur	tur	crh*
rank – distance		0.7	6	0.7	7
label		Turkish	Türkiye Türkçesi	turkish	Turkish, Crimean

Table 3: Excerpt output of language mapping. * indicates mismatches.

Resource	Label Accuracy	Instance Accuracy
SIL *dice coefficient*	81%	99.50%
SIL *levenshtein*	72%	99.42%
BabelNet *dice coefficient*	**91%**	99.87%
BabelNet *levenshtein*	**89%**	99.85%
SIL + BabelNet		
dice coefficient	**91%**	99.87%
levenshtein	**89%**	99.85%

Table 4: Accuracy of language mappings

Net is more accurate in matching language labels, largely because it contains translations, synonyms and obsolete spellings of most, even rare or dialectal, languages. SIL on the other hand only contains the English representation of each ISO code, failing to induce certain mappings. Furthermore, the Dice coefficient string similarity algorithm yields more accurate results compared to the Levenshtein distance metric. We hypothesize that this is mainly due to the fact that the Dice coefficient is more lenient compared to the Levensthein metric as it is insensitive to the order of words. For instance, using Dice coefficient, the input label 'Quechua de Cotahuasi (Arequipa)' matches 'Cotahuasi Quechua' correctly. With the Levenshtein algorithm, however, using the same input as earlier, the label 'Quechua cajamarquino' is mistakenly matched instead.

Overall, combining BabelNet and SIL yields the same normalization accuracy as BabelNet alone.

Resource	Duplicate Titles	Duplicate URLs
CLARIN (same contributing institute)	50,589	20
Datahub.io	0	55
META-SHARE	63	967

Table 5: The number of intra-repository duplicate labels and URLs for resources

Nonetheless, we can observe a slight decrease in the average distance returned by the Levensthein algorithm. The addition of a multilingual semantic database, such as BabelNet, positively affects the ability to match obsolete names in different languages.

4.5 Type

The type property is used primarily to describe the kind of resource being described. For META-SHARE, we can rely on the structure of resources to extract one of four primary resource types, namely, 'Corpus', 'Lexical Conceptual Resource', 'Lexical Description' and 'Tool Service'. However, for the other sources considered in this paper the type field permits free text input. In order to enable users to query resources by type we ran the Babelfy entity linking algorithm (Moro et al., 2014) to identify entities in the string and then manually selected elements from this list of entities that described the kind of resource, such as 'corpus'. In this way we extracted, 143 categories for language resources while still ensuring that syntactic variations were accounted for. The top 10 categories extracted in this way were: 'Sound', 'Corpus', 'Lexicon', 'Tool' (software), 'Instrumental Music'[9], 'Service', 'Ontology', 'Evaluation', 'Terminology' and 'Translation software'.

5 Duplicate detection

As we are collecting and indexing metadata records from different repositories, it is possible to find duplicates, that is records that describe the same actual resource. In fact, duplicate entries did not only occur across repositories (we dub these *inter-repository duplicates*) but also within the same resource (referred to as *intra-repository duplicates*). We expand the definition of inter-repository by noting that CLARIN is sourced from a number of different contribut-

[9]These resources are in fact recordings of singing in under-resourced languages

ing institutes and there are duplicates between institutes, thus we consider links between records of different CLARIN institutes as *inter-repository*. Similarly, there has been no attempt to manage duplicates in LRE-Map and so we handle all links between LRE-Map records as *inter-repository*.

In order to detect duplicates, we rely on two properties that should be unique across entries, that is the title and the 'access URL'. In Table 5 we show the number of records with duplicate titles or URLs. Manual inspection of these duplicates yielded the following observations:

META-SHARE META-SHARE contains a number of duplicate titles. However, these title duplicates seem to be errors in the export and can thus be easily corrected.

CLARIN Many resources in CLARIN are described across many records. For example, in CLARIN there may be one different metadata record for each chapter of a book or recording within an audio or television collection, or in at least one case ("The Universal Declaration of Human Rights") a record exists for each language the resource is available in. Thus, we decided to merge the entries which share the same title and same contributing institute in CLARIN.

Datahub.io The creation method of DataHub prevents the creation of different entries with the same title, so duplicate titles do not occur in the data. However, we found a number of entries having the same download URL. This is due to the fact that different resources share SPARQL endpoints or download pages, but the records did not describe the same resource and so we did not merge these resources.

Table 6 shows the number of resources with the same title (Duplicate Titles), same URL (Duplicate URLs) as well as same title **and** same URL within and across repositories. We apply the following strategy in handling duplicates:

Intra-repository duplicates As intra-repository duplicates are mostly either system errors or series of closely related resources, we simply merge the corresponding metadata entries. If a property is one-to-one we take only the first value.

Duplication	Correct	Unclear	Incorrect
Titles	86	6	8
URLs	95	2	3
Both	99	1	0

Table 7: Precision of matching strategies from a sample of 100

Property	Record Count (As percentage of all records)	Triples
Access URL	91,615 (91.6%)	191,006
Language	50,781 (50.7%)	98,267
Type	15,241 (15.2%)	17,894
Rights	3,080 (3.0%)	8915
Usage	3,397 (3.4%)	4,530

Table 8: Number of records and facts harmonized by our methods

Inter-repository duplicates Inter-repository duplicates represent multiple records of the same underlying resource, they are linked to one another by the 'close match' property.

Note we do not remove duplicates from the dataset we either combine them into a more structured record or mark them as deprecated.

We evaluate the precision of this approach on a sample of 100 inter-repository entries identified as duplicates according to the above mentioned approach. We manually classify the matches into *correct*, *incorrect* as well as *unclear*, if there was insufficient information to make a decision, the resources overlapped or were different versions of each other. Table 7 shows these results. We see that with 99% precision the method identifying duplicates if both title and URL match yields the best results. While the recall is difficult to assess, an analysis of the data quickly reveals that there are many duplicates not detected using this method. For example, for the Stanford Parser (De Marneffe et al., 2006), we find metadata records with all of the following titles: "Stanford Parser", "Stanford Dependency Parser", "Stanford Lexicalized Parser", "Stanford's NLP Parser", "The Stanford Parser", "The Stanford Parser: A Lexicalized Parser".

6 Discussion

The rapid developments of natural language processing technologies in the last few years has resulted in a very large number of language resources being created and made available on the

Resource	Resource	Duplicate Titles	Duplicate URLs	Both
CLARIN	CLARIN (other contributing institute)	1,202	2,884	0
CLARIN	Datahub.io	1	0	0
CLARIN	LRE-Map	72	64	0
CLARIN	META-SHARE	1,204	1,228	28
Datahub.io	LRE-Map	59	5	0
Datahub.io	META-SHARE	3	0	0
LRE-Map	LRE-Map	763	454	359
LRE-Map	META-SHARE	91	51	0
All	All	3,395	4,686	387

Table 6: Number of duplicate inter-repository records by type

web. In order to enable these resources to be reused appropriately it is necessary to properly document resources and make this available as structured, queriable metadata on the Web. Current approaches to metadata collection are either *curatorial*, where dedicated workers maintain metadata of high quality, such as the approach employed by META-SHARE. This approach ensures metadata quality but is very expensive and as such it is unlikely that it will be able to handle the vast number of resources published every year. In contrast, *crowd-sourced* resources rely primarily on self-reporting of metadata, and this approach has a high recall but is very error-prone and this unreliability can be plainly seen in resources such as LRE-Map. In this paper, we have aimed to break this dichotomy by aggregating resources from both curated and crowd-sourced resources, and applied natural language processing techniques to provide a basic level of compliance among these metadata records, and have achieved this for a large number of records as summarized in table 8. In this sense we have considered a small set of properties that we regard as essential for the description and discovery of relevant language resource, that is: resource type, language, intended use, and licensing conditions. For the language property we have shown that it can be harmonized across repositories with high accuracy by mapping values to a controlled vocabulary list, although the data indicated that there were still many languages which were not covered in the ISO lists. For the type, rights and usage properties, whose content is not as limited, it is harder to harmonize but we were still able to show that in many cases these results can be connected to known lists of values. This is important as it would allow for easier queries of the resource.

Besides harmonizing values of data, we see two further key aspects to ensure quality of the metadata. First, broken links should be avoided as they are indicators of low curation and low quality. Thus, we automatically detect such broken URLs and remove them from the dataset. A second crucial issue is the removal of duplicates, which are also a sign of low quality.

We have investigated different strategies for detecting duplicates. We observed that the case in which two metadata records have been provided to different repositories is common. When integrating data from different repositories, these entries become duplicated. In other cases, particularly in CLARIN, different metadata records are created for parts of a resource. Genuine duplication likely affects about 7% of records, underlining the value of collecting resources from multiple sources. We further note that it is important to take a high precision approach to deduplication as the merging of non-duplicate resources can hide resources entirely from the query. Thus, we have proposed high-precision methods for detecting such duplicates.

Finally, we note that the data resulting from this process is available under the Creative Commons Attribution Non-Commercial ShareALike License and the data can be queried through a portal, which is available at **URL anonymized**. Furthermore, all code described in this paper is accessible from a popular open source repository.[10]

7 Conclusion

We have studied the task of harmonizing records of language resources that are heterogeneous on several levels and have shown that the applica-

[10]To remain anonymous we cannot include URLs for these resources at this point

tion of NLP techniques allows to provide common metadata that will better enable users to find language resources for their specific applications. We note that this work is still on-going and should be improved in not only the accuracy and coverage of harmonization, but also in the number of properties that are harmonized (authorship and subject topic are planned). We hope that this new approach to handling language resource metadata will better enable users to find language resources and assist in the creation of new domains of study in computational linguistics.

Acknowledgments

LingHub was made possible due to significant help from a large number of people, in particular we would like to thank the following people: Benjamin Siemoneit (Bielefeld University), Tiziano Flati (University of Rome, La Sapienza), Martin Brümmer (University of Leipzig), Sebastian Hellmann (University of Leipzig), Bettina Klimek (University of Leipzig), Penny Labropolou (IEA-ILSP), Juli Bakagianni (IEA-ILSP), Stelios Piperidis (IEA-ILSP), Nicoletta Calzolari (ILC-CNR), Riccardo del Gratta (ILC-CNR), Marta Villegas (Pompeu Fabra), Núria Bel (Pompeu Fabra), Asunción Gómez-Pérez (Universidad Politécnica de Madrid) and Christian Chiarcos (Goethe-University Frankfurt).

This work was funded by the LIDER ("Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe"), an FP7 project refererence number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.

References

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry-and component-based metadata framework. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 43–47.

Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, page 1.

Christopher Brooks and Gord McCalla. 2006. Towards flexible learning object metadata. *Continuing Engineering Education and Lifelong Learning*, 16(1/2):50–63.

Elena Cabrio, Alessio Palmero Aprosio, and Serena Villata. 2014. These are your rights: A natural language processing approach to automated RDF licenses generation. In *The Semantic Web: Trends and Challenges*, pages 255–269. Springer.

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo, and Claudia Soria. 2012. The LRE Map. Harmonising community descriptions of resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1084–1089.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012. The Open Linguistics Working Group of the Open Knowledge Foundation. In *Linked Data in Linguistics*, pages 153–160. Springer.

Khalid Choukri, Victoria Arranz, Olivier Hamon, and Jungyeul Park. 2012. Using the international standard language resource number: Practical and technical aspects. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 50–54.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Christian Federmann, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schröder. 2012. META-SHARE v2: An open network of repositories for language resources including data and tools. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3300–3303.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Harris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, et al. 2012. The META-SHARE metadata schema for the description of language resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1090–1097.

Michael Khoo and Catherine Hall. 2010. Merging metadata: a sociotechnical study of crosswalking and interoperability. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 361–364. ACM.

Joseph Mariani, Christopher Cieri, Gil Francopoulou, Patrick Paroubek, and Marine Delaborde. 2014. Facing the identification problem in language-related scientific data analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 2199–2205.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Mikael Nilsson. 2010. *From interoperability to harmonization in metadata standardization*. Ph.D. thesis, Royal Institute of Technology.

Javier Nogueras-Iso, F Javier Zarazaga-Soria, Javier Lacasta, Rubén Béjar, and Pedro R Muro-Medrano. 2004. Metadata standard interoperability: application in the geographic information domain. *Computers, environment and urban systems*, 28(6):611–634.

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 36–42.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Victor Rodriguez-Doncel, Serena Villata, and Asuncion Gomez-Perez. 2014. A dataset of RDF licenses. In Rinke Hoekstra, editor, *Proceedings of the 27th International Conference on Legal Knowledge and Information System*, pages 187–189.

Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

Herbert van de Sompel, Michael L Nelson, Carl Lagoze, and Simeon Warner. 2004. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10(12).