Converting linguistic datasets into interoperable Linked Data resources

The case of WALS, IDS and WOLD

Martin Brümmer
Universität Leipzig, Institut für Informatik, AKSW, E-mail: bruemmer@informatik.uni-leipzig.de

Abstract. In this paper I describe the conversion of three linguistic datasets, namely WALS, IDS and WOLD, into RDF. I focus my discussion on the challenges encountered in transforming detailed linguistic datasets, and in particular their disparate internal structures and semantic contents, into interoperable Linked Data. I then test the syntactic and semantic interoperability achieved by linking these resources together in the LLOD and I highlight the general problems involved in making broad cross-linguistic data sources interoperable.

Keywords: multilingual, dictionary, language structure, Linked Open Data, RDF, interoperability

1. Introduction

In this paper I describe the pitfalls involved in converting three datasets from different linguistic domains into an interoperable Linked Data resource and I show how to convert these resources according to current best practices. The datasets include the World Atlas of Language Structures (WALS, [5]),[1] the Intercontinental Dictionary Series (IDS, [4])[2] and the World Loanword Database (WOLD, [6])[3]. Each dataset represents a novel and important body of linguistic research and all have been developed by the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology (MPI EVA). Each resource is implemented in an online-accessible and dataset-specific database, because each dataset has a different research focus and is therefore different in scope and internal structure. In this paper my aim is to highlight how to convert existing (and future) datasets into an interoperable format that leverages Semantic Web technologies to ensure the proper conversion of linguistic data into the Linguistic Linked Open Data cloud.[4]

This paper is structured as follows. In Sections 2-4, I describe the content and structure of each dataset. Then in Section 5 I highlight issues of general interoperability and focus on challenges specific to each dataset. Specifically, my aim is to address issues of interlinking these resources with the LLOD and on how to undertake vocabulary design. Finally, in Section 6 I present my conclusions.

2. World Atlas of Language Structures (WALS)

As described by its editors, WALS is: “a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors”[5]. This dataset contains 192 structural typological *features* from a broad sample of the world’s languages and each feature is categorized into one of 144 different chapters. For example, there are chapters describing phenomena such as, the number of gen-

[1] http://wals.info/
[2] http://lingweb.eva.mpg.de/ids/
[3] http://wold.livingsources.org/
[4] http://linguistics.okfn.org/resources/llod/

ders found in languages or the existence of different tenses and word orders across languages. Each feature contains a number of possible *values*. For example, the chapter on word order, titled “Order of Subject, Object and Verb”, describes the canonical word order found across languages, using the feature values “SOV”, “SVO”, “VSO”, etc.

In total, WALS has data on 2678 languages. Most chapters contain about 500 data points, i.e., tuples of the type `(language,feature)`, where each language in the chapter contains a language-specific value. Furthermore, the data includes language names, alternative language names, ISO 639-3 codes, geographical coordinates of where each language is spoken, and a genealogical classification of each language into language family and genera.

The conversion of the WALS dataset was based on a MySQL-database dump.[5] I developed a new vocabulary to convert the main data; my data model is illustrated in Figure 1. The URI formats and prefixes are specified in Table 1. Although I used existing vocabularies such as DCTERMS[6] and WGS84 Geo Positioning[7], most classes and properties had to be newly defined, due to the lack of an existing ontology that was granular enough to describe the WALS dataset. Each language resource is uniquely identified by a URI containing a three letter WALS code, i.e., an internal unique identifier developed by the WALS team. I could have used ISO 639-3 codes for this purpose, but the identification of languages in WALS was made before the ISO 639-3 standard was established and there is not always a 1:1 mapping between WALS code and ISO 639-3 code[5]. Instead, existing ISO 639-3 codes were used to link the WALS data to a number of different language resources, including Lexvo,[8] Glottolog/Langdoc,[9] and SIL[10]. The WALS features are modeled as properties, which connect the languages to data points that contain the provenance of the feature information as a literal in a `dcterms:references` element. The feature values are linked to the data points via the `wals:hasValue` property. The value information is contained in the `rdfs:label` and the `dcterms:description` elements. For example, the value labeled “SOV” has the literal “Subject-object-verb (SOV)” in its `dcterms:description` element. The problem, however, is that all values contain this kind of information as literals. Ultimately, there is no well-defined structure due to the heterogeneous nature of the different value types. Thus, the information can be read and understood by humans, but is not easily parsed by computers. For example, the feature of word order is useful for automatically correcting grammar, but it cannot be programmatically learned from the WALS data without knowing the specific structure of the feature text. Therefore, WALS is a huge collection of information useable by linguists, but the simple RDF structure introduced in this paper does not yet make it usable knowledge in a computational sense.

The RDF conversion of WALS contains 499,112 triples and over 6000 links to additional language resources. It is published online as a dump[11]and can be queried via `http://mlode.nlp2rdf.org/sparql`.

3. The Intercontinental Dictionary Series (IDS)

The IDS is a large collaborative effort, compiled by editors from a number of international sources. It is a multilingual dictionary organized in chapters by semantic category to allow easy comparison of lexical items across different languages [4]. IDS contains 23 *chapters*, including categories such as “the physical world”, “animals” or “food and drink”. These chapters include 1310 language-specific *concepts* with reference translations in English, French, Russian, Spanish and Portuguese. Associated with these concepts are the actual language-specific lexical items, called *entries*, of which there are 280,000, spanning 215 languages.

For the conversion of the IDS data into RDF, I was given access to a PostgreSQL database dump of the data. As a first step I analyzed the dataset and produced the data model illustrated in Figure 2. Table 2 shows the URI formats and prefixes.

To ensure interoperability, it was important that tables containing language data were merged in the language class. The language class contains additional information, including ISO 639-3 codes, official (ISO) names of the languages, alternative language names,

[5] The data can also be downloaded as CSV files: `http://wals.info/export`

[6] `http://dublincore.org/documents/dcmi-terms/`

[7] `http://www.w3.org/2003/01/geo/wgs84_pos#`

[8] `http://www.lexvo.org/`

[9] `http://www.glottolog.org/`

[10] `http://sil.org/`

[11] `http://mlode.nlp2rdf.org/datasets/wold.nt.gz`

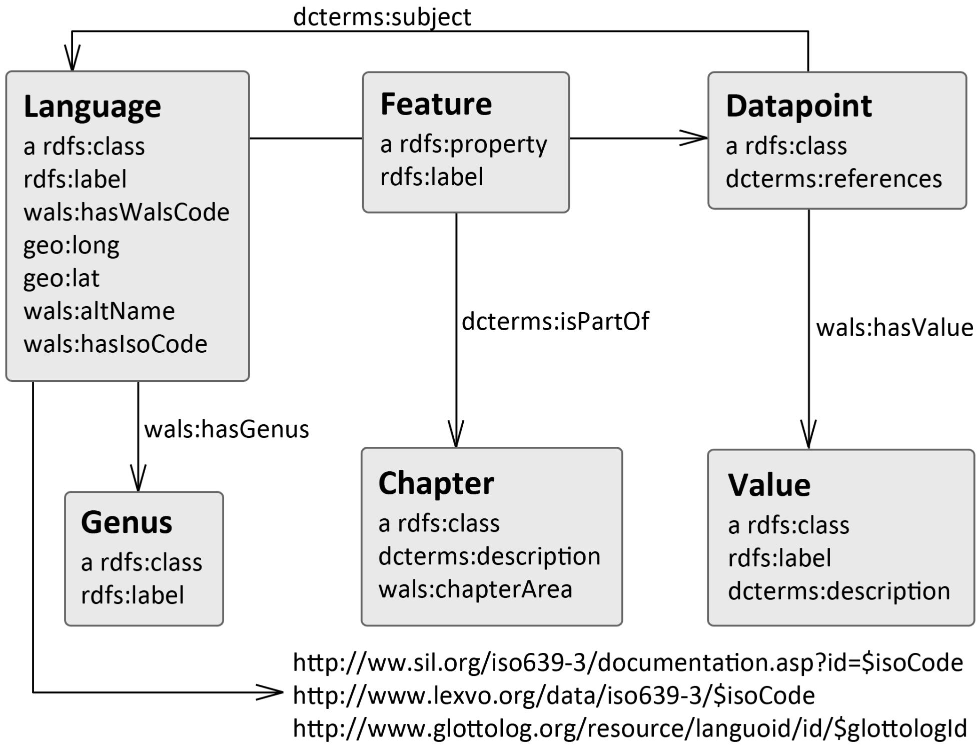

Fig. 1. WALS data model diagram

Table 1
WALS URI formats

Class	URI format
Language	`http://mlode.nlp2rdf.org/resource/wals/language/$languageId`
Feature	`http://mlode.nlp2rdf.org/resource/wals/feature/f$featureId`
Datapoint	`http://mlode.nlp2rdf.org/resource/wals/datapoint/$languageId-f$featureId`
Genus	`http://mlode.nlp2rdf.org/resource/wals/genus/$genusId`
Chapter	`http://mlode.nlp2rdf.org/resource/wals/chapter/ch$chapterId`
Value	`http://mlode.nlp2rdf.org/resource/wals/value/f$featureId-$valueId`
Prefix wals:	`http://mlode.nlp2rdf.org/resource/wals/vocabulary/`

and the source information of the data. The data provenance information is particularly important with regard to IDS, because, as mentioned, it is an international collaborative effort. Data entry, compilation, consultation and sources are also converted into RDF, each as a literal in its respective field. Although the information is thereby contained in the RDF files, the apparent problem with this procedure is, that they are not resources on their own, which would be desirable for making granular queries. Language resources are ordered into larger classes, with the `inLangClass` property. Theses classes serve as an unspecific amalgamation of different kinds of language groups. For example, an entry would be part of a dialect, which in turn would be part of an ethnolect as the corresponding language class. The ethnolect would then be part of a larger language family or genus. A hierarchy of subclasses is formed: Each Langclass is linked to a class of higher level by the `skos:broader` and an additional `dcterms:relation` property. Due to the granularity of these language classes and the lack of additional data, it was not possible to further distinguish them into dialects, ethnolects, families or genera, like in the case of WALS.

The IDS entries are divided into two kinds:

(a) The reference translations, which contain the ids:$translation properties. Their label contains the English translation of the entry. They are furthermore linked to DBpedia Wiktionary.[12] IDS lacks part-of-speech information, therefore the linking approach described in [8] could not be used. The only check performed was for the existence of the Wiktionary resource of the string representing the word form. Because of this procedure, the correct word resource will be linked in many cases, but correctness can not be assured.

(b) The lexical entries, which contain rdfs:label. The label is *not always normalized* and may contain multiple forms, separated by semicolons, parenthesis or one or more minus signs. The entries may also contain further data like alternative forms or additional inflections of the entry. This kind of entry is also linked to the relevant language resource by the gold:inLanguage property and to the entry containing the reference translations via dcterms:relation.

Both types of entries are linked to the relevant chapter via dcterms:isPartOf.

To confirm to Linked Data principles, the language resources were again linked to WALS, Glottolog/Langdoc and Lexvo. Furthermore, the chapter resources are linked to WOLD semantic fields, because the latter are based on the former.

The resulting RDF dataset contains 1,984,321 triples and 216 links to each WALS, Glottolog and Lexvo. There are additional 1832 links to DBPedia Wiktionary for the lexical entries and 22 links to WOLD semantic fields. It is published online as a dump[13]and can be queried via http://mlode.nlp2rdf.org/sparql.

4. The World Loanword Database (WOLD)

WOLD is a database of vocabularies of about 2000 entries for 41 languages. Each entry includes information about its loanword status, source words and associated meanings. This resource provides the ability to find out about donor languages, and by means of geographical coordinates, the geographical distributions of the characteristics of loanwords. In WOLD, the donor languages are not necessarily part of the 41 languages analyzed, so WOLD provides a varying amount of information for a total of 395 languages, some of them only mentioned by name. WOLD was published in 2009 under the Creative Commons Attribution 3.0 License.[14] An RDF+XML Version was added in 2010 as Linked Data and is available from the WOLD website. There are 837,828 triples and around 17000 links to DBPedia, WALS and SIL. There is no SPARQL-Endpoint on the project page available, so it was established at http://mlode.nlp2rdf.org/sparql. An RDF dump was compiled and published.[15]

Although WOLD was not converted into RDF for this paper, I will describe its structure nevertheless to highlight certain problems common to converting datasets into Linked Data and linking them to the LLOD. A description of WOLD's RDF data is also beneficial in describing how it is linked to the Linked Data survey in this paper, as will become clear in Section 5.3.

The RDF+XML version of the WOLD contains seven classes and uses existing ontologies, including GOLD,[16] the WordNet 2.0 schema,[17] and SKOS.[18] The WOLD language resources contain only the name of the language and geographical coordinates as kml:coordinates.[19] ISO 639-3 language codes are not explicitly defined in the data. However, they are contained in links to the more widespread languages in SIL. These links exist for about 62% of the languages in the WOLD dataset, thus limiting the interoperability of this dataset at the level of language code. The same applies to links to WALS. Although language families and genera are also linked to WALS in the WOLD HTML version, there are no such links in the RDF serialization. Recipient languages which borrowed words from donor languages can also be found in the HTML representation but not in the RDF. These are issues that need to be resolved to make the RDF Linked Data version of WOLD more interoperable with the other MPI EVA datasets and with the LLOD in general.

[12] http://dbpedia.wiktionary.org/resource/

[13] http://mlode.nlp2rdf.org/datasets/ids.nt.gz

[14] http://creativecommons.org/licenses/by/3.0/de/

[15] http://mlode.nlp2rdf.org/datasets/wold.nt.gz

[16] http://linguistics-ontology.org/gold

[17] http://www.w3.org/2006/03/wn/wn20/

[18] http://www.w3.org/2009/08/skos-reference/skos

[19] http://www.opengis.net/kml/2.2

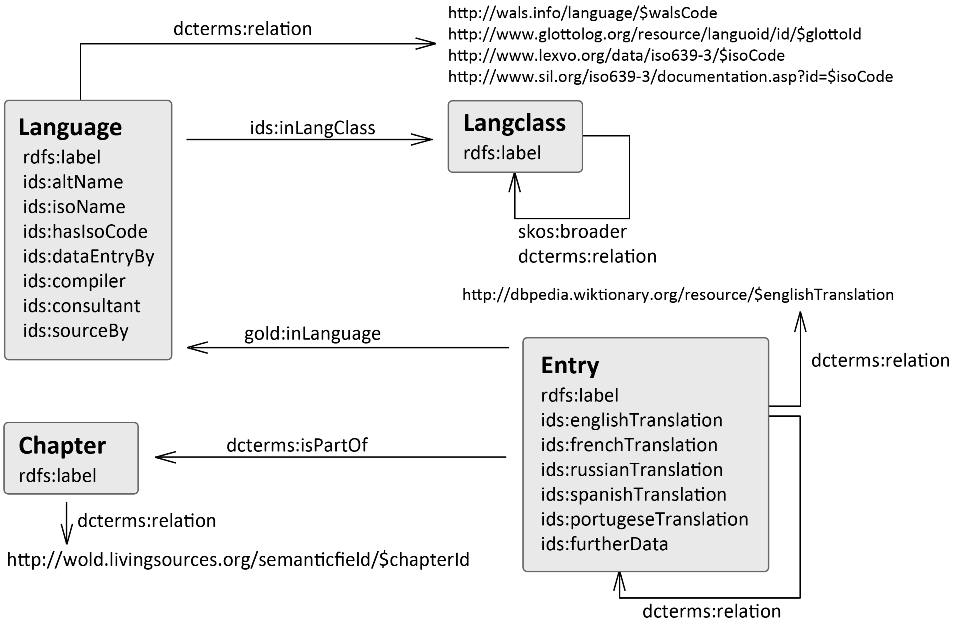

Fig. 2. IDS data model diagram

Table 2
IDS URI formats

Class	URI format
Language	`http://mlode.nlp2rdf.org/resource/ids/language/$languageId`
Langclass	`http://mlode.nlp2rdf.org/resource/ids/langclass/$langclassId`
Entry	`http://mlode.nlp2rdf.org/resource/ids/entry/$entryId`
Chapter	`http://mlode.nlp2rdf.org/resource/ids/chapter/$chapterId`
Prefix ids:	`http://mlode.nlp2rdf.org/resource/ids/vocabulary/`

Language resources link vocabularies, which consist basically of a number of `dcterms:hasPart` properties that link resources about words. The latter provide an orthographic representation, a lexical form and links to possible source words. The HTML representation of WOLD again provides more information, including the part of speech, comments, additional references, and a label for borrowed status that includes assessments about a word's likely source. Meanings are linked by words in a peculiar form: The `wn:sense` property links to a WordSense resource without its own identifier. Instead the URI is derived from a meaning resource and the identifier of the word is attached as a fragment. So the `wn:sense` property of a word `http://wold.livingsources.org/word/$wordId.rdf` links the resource `http://wold.livingsources.org/meaning/$meaningId#$wordId`, which is not available as RDF. To retrieve the RDF representation, one has to retrieve the meaning resource without the fragment `http://wold.livingsources.org/meaning/$meaningId.rdf` to find the meaning as well as all contained word sense resources. This does not adhere to Linked Data principles and it is limiting interoperability. Meanings are ordered into semantic fields, a broader semantic classification based on IDS chapters. Semantic fields are equivalent to IDS chapters by name and scope.

In general, the RDF+XML serialization of WOLD is lacking in terms of volume of data converted. The HTML pages show a bigger picture which cannot yet be accessed as Linked Data.

5. Interoperability

Interoperability as defined by [7] is divided into syntactic and semantic interoperability. RDF as a data for-

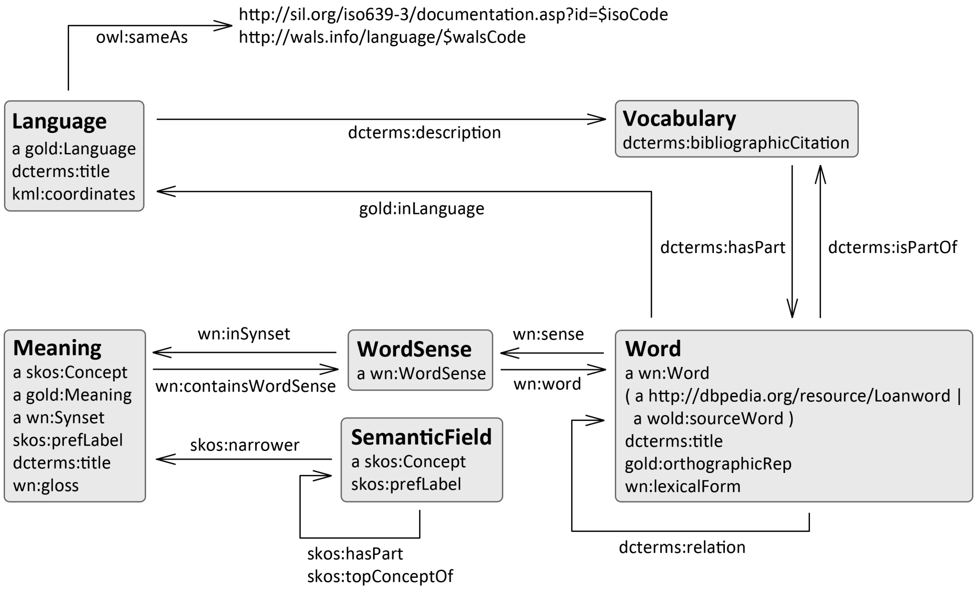

Fig. 3. WOLD data model diagram

Table 3
WOLD URI formats

Class	URI format
Language	`http://wold.livingsources.org/language/$languageId`
Vocabulary	`http://wold.livingsources.org/vocabulary/$vocabularyId`
Word	`http://wold.livingsources.org/word/$wordId`
Meaning	`http://wold.livingsources.org/meaning/$meaningId`
WordSense	`http://wold.livingsources.org/meaning/$meaningId#$wordId`
SemanticField	`http://wold.livingsources.org/semanticfield/$fieldId`
Prefix gold:	`http://purl.org/linguistics/gold/`
Prefix wn:	`http://www.w3.org/2006/03/wn/wn20/schema/`

mat provides syntactic interoperability. The standard is also well-defined and can be processed by a number of existing tools. This is an important benefit, as established frameworks like Jena[20] allow easy and granular data access, aggregation and manipulation. Web tools like OntoWiki[21] can also be used to enhance collaborative research, as shown in [9]. On the other hand, semantic interoperability, as a means of “consistent interpretation of exchanged data” [7, p. 2], is dependent on common definitions of concepts in an ontology or vocabulary. If the vocabularies used to describe the linguistic data do not intersect, consistent interpretation is not possible. Usual routines of vocabulary creation during dataset conversion must therefore be examined and adapted to the linguistic domain. In Section 5.1 I illustrate this in more detail.

Another hurdle for interoperability is the interlinking of different datasets, highlighted in Section 5.2. Although one could argue that by linking metadata in RDF conceptual interoperability is automatically given, e.g. [3], but two major prerequisites would be ignored:

(a) The concept itself, i.e., the property of the resulting triple, has to be well-defined. In the linguistic domain, this is dependent on scientific consensus, which is hard to come by, particularly in specialized linguistic subdomains. Although efforts are made to elevate the issue by mapping different annotation schemes [1,2], or domain specific concepts [10], the width of the linguistic domain makes this approach difficult and work-intensive.

[20] `http://jena.apache.org/`
[21] `http://ontowiki.net/`

In the biomedical domain, this has been alleviated by the creation of central hubs for working with ontologies, such as the BioPortal.[22] A hub like this can provide access to existing ontologies of the domain, as well as mappings between these ontologies. However, such hubs are missing in linguistics.

(b) If the concept is well-defined, it is done by textual descriptions, often in `rdfs:comment` elements, which are not semantically interpretable by machines, and this hinders interoperability. Furthermore, this definition must be clear to the compiler as well as the user of the data; this requires clear documentation. An especially pressing example is the interlinking of resources describing languages themselves.

In spite of these problems, Linked Open Data has some advantages regarding interoperability. Opening so-called *data silos*, such as dataset-specific databases, to the public via the Semantic Web enhances interoperability and the scope of language resources, but this does not guarantee semantic interoperability. Therefore, even with structural interoperability provided by RDF standards, without ontologies mighty enough to capture at least essential parts of linguistic subdomains, semantic interoperability is not achieved.

5.1. Vocabularies

To grant automatic interpretation or semantically correct integration of different datasets, the vocabularies and ontologies used must be compatible. At the time of writing, there is no commonly accepted and complete vocabulary for the description of linguistic resources. The breadth of the linguistic domain may be the biggest hurdle in achieving such a vocabulary, but a distinct problem lies in the ad-hoc definition of vocabularies in the course of dataset conversion. Although the use of GOLD, WordNet 2.0 schema, SKOS, OWL[23] and DCTERMS is encouraged and widespread, vocabularies are often defined bottom-up during the conversion into RDF. Furthermore, these definitions are made by Linked Data experts unfamiliar to the specific domain, like in the case of IDS in this paper. The reasoning behind this practice is to avoid the definition of semantically inappropriate descriptors. This problem is typical of the interdisciplinary use of Linked Data, e.g. [9], and must be addressed in the future for resources being linked in the LLOD. In the mean time, efforts like OLiA [1,2] and ISOcat [11] help to tackle the issue by mapping different ontologies and data categories.

5.2. Linking of Language Resources

The most basic concept in the domain of Linguistic Linked Open Data is the concept of *language*. While different datasets focus on specific subdomains of linguistic research (and interoperability of datasets may often not be possible on these specific levels), the languages that the datasets describe can still be interlinked with language resources from other datasets. The biggest problem to solve in this regard is language identification. If there were generally accepted language identifiers, this would not be a problem. Thus the ISO 639-3 language name identifier was introduced for this specific purpose and it has gain wide acceptance within the linguistics community.

ISO 639-3 aims to provide language name identifiers for every language, including living, extinct and ancient languages. It is being expanded to include dialects as well [12]. This standard contains codes for over 7700 languages. While this number should be sufficient for most linguistic purposes, there still exists a problem in data compilation. Linguists may disagree about the specific boundary of small dialects, ancient languages, or even word forms of hard-to-specify origin. In Linked Data, this problem can be addressed by the use of specific properties. For example, the `owl:sameAs` property should not necessarily be used to link languages to other resources with the same ISO code. The property suggests semantic equivalence, which overrides subtle differences of the definition of language in different datasets. Rather, as shown by the conversion of WALS and IDS in this paper, a *relation* property such as `dcterms:relation` should be used.

Another important question is: What datasets should be linked? In the LLOD, a number of datasets have been established as reference datasets for this purpose, including Glottolog/Langdoc, Lexvo and WALS. Lexvo can easily be linked, because its URIs contain ISO 639-3 codes. For Glottolog and WALS, web services are in the process of being established, to map the ISO 693-3 codes to the specific identifiers of these datasets. Such a web service would be especially interesting for Glottolog, as its stated goal is mapping all kinds of language varieties to a unique ID. However,

[22] `http://bioportal.bioontology.org/`

[23] `http://www.w3.org/2002/07/owl#`

to reliably map a language resource to a Glottolog ID, the service would not only need to map ISO codes, but also the names of language resources and maybe even the direct hierarchical neighborhood. This would be due, as one can not reliably tell from the data, on what grounds an ISO code was assigned to a language. It may be the code of the broader class of the language, for example. For this paper, links to Glottolog where established by using a static, intellectually compiled list of WALS, ISO and Glottolog IDs. Bear in mind that this suffers from the same potential error, but was disregarded because of the overall high quality of the MPI datasets.

5.3. Practical interoperability on the example of the described datasets

To expand on the aforementioned challenges, interoperability was tested on the datasets described in this paper. Each dataset was loaded into a triple store and then queried with SPARQL to achieve different goals, as specified by the queries found in the appendices. The first objective was matching languages to integrate information about different languages from the three different sources. This objective was done by comparing ISO 639-3 codes, which proved to be a hurdle, as explained in 5.2. WOLD contains ISO 639-3 codes for only 62% of its language resources, severely limiting the possibilities of interlinking WOLD with other Linked Data sources in the LLOD. Further, the ISO 639-3 codes are not directly specified as literals, but as parts of URIs, which requires string processing in the SPARQL-queries to extract them.

Table 4 shows the results for the language matching test. The number of triples refers to the output of the SPARQL queries. It is usually higher than the number of matched languages, as language resources of the datasets often have more than one ISO 639-3 code. Percentages shown refer to the percentage of languages of the respective dataset that matched to the languages of the other datasets. For WOLD, the percentage of languages containing ISO 639-3 codes was noted separately, to reflect that the quantity of results is dependent on the existence of these codes.

The best results were obtained for matching WOLD and IDS languages to WALS. The positive results are due to the the high number of languages contained in WALS. 78% of WOLD language resources with ISO 639-3 codes and 83% of IDS language resources can therefore be enriched with WALS information, including geographical coordinates and language feature information. Conversely, only 7-8% of WALS language resources can profit from the other datasets. A special case may be the comparison of geographical coordinates of WALS and WOLD, although they are annotated differently.

Around one fifth of the available language resources from WOLD and IDS can be matched. WOLD's 41 languages achieved 100% coverage. Integrating both datasets could yield interesting results because IDS is a multilingual dictionary and WOLD contains loanword and source word information, thus combining the data from both datasets is attractive to researchers. Due to the nature of WOLD, the chapters of IDS and the semantic fields of WOLD already match. To find out how big of a hindrance the relatively small amount of matched languages is to the combination of the datasets, the triple store was again queried via SPARQL, but this time to match on the level of individual words as well as languages. One has to remember that IDS entries are not normalized and often contain multiple lexical forms and other information. Therefore for the query I used a `FILTER(regex(?idsEntry,?woldWord))` expression, which searches for occurrences of the WOLD word string in the IDS entry label. This is not very precise, but in this case it was the only way to find a sufficient number of matches. Although only 18% of the WOLD languages could be linked to IDS languages, this query yielded 13,030 triples. Out of a total number of 57,926 WOLD words, this means that 22% of WOLD words can be linked to similar IDS entries, which, via the linked reference entries, provide translations into English, French, Russian, Spanish and Portuguese. On the other hand, out of a total of 282,671 IDS entries, only 1% of them can be enhanced with loanword and source word information from WOLD. Although I am sure that these results can be improved by more exhaustive linking of language resources, there is currently no way of testing this hypothesis.

Matching languages over all three datasets did not yield many results, but those resources are likely to be among the bigger and more thoroughly researched languages, as they feature ISO 639-3 codes in all examined datasets. Again, all 41 languages focused on by WOLD were among the results.

Although the datasets were converted largely without interoperability as a specific aim, the results of the queries show interesting possibilities for future data integration. If the Linked Data version of WOLD contained all the information found in the HTML repre-

Table 4
Matched languages and percentage of coverage per dataset

Datasets	Number of triples	Languages per dataset and coverage of matching		
WOLD+WALS	282	WOLD	WALS	
		190 (48%, 78% iso)	218 (8%)	
IDS+WALS	225	IDS	WALS	
		179 (83%)	197 (7%)	
IDS+WOLD	57	IDS	WOLD	
		49 (23%)	44 (11%, 18% iso)	
IDS+WALS+WOLD	83	IDS	WALS	WOLD
		46 (21%)	53 (2%)	41 (10%, 17% iso)

sentation, integration with the IDS could be further enhanced, or at least it could be done more precisely. The part of speech would be particularly interesting. The points made in Section 5.2 cannot yet be seen as proven, as I cannot compare how many data will be retrieved if more languages were linked. While the relatively small number of matches between the smaller datasets of WOLD and IDS can be explained by the different focuses of these projects, the greater coverage of WOLD languages with ISO codes shows the importance of incorporating commonly accepted language name identifiers such as ISO 639-3 codes or Glottolog IDs into the datasets.

6. Conclusion

The datasets described in this paper present a useful contribution to the LLOD-Cloud. The datasets' precision and the structure of the original data turned out to be an excellent starting point for the RDF conversion. Due to their common provenance, they were especially suited for an examination of dataset interoperability, which showed the advantages as well as the shortcomings in the datasets and the LLOD-cloud as a whole. ISO 639-3 codes proved to be essential for interlinking resources by languages and for ensuring cross-dataset compatibility. The practical part of the paper has shown that size-able parts of different datasets can be easily enriched with data from other sources, even if the findings still have to be examined for correctness by researchers or, in some cases, algorithms. It was also shown that semantic interoperability will be an important issue for future research. Nevertheless, the advantages of RDF as a data format for achieving syntactic interoperability outweigh the disadvantages of using RDF and Linked Data.

7. Acknowledgements

The conversion of the WALS dataset was supported by Sebastian Nordhoff, providing major help to develop the vocabulary. The WOLD RDF serialization was developed by Robert Forkel. Further thanks go to Sebastian Hellmann, for helping with technical issues and providing the platform of the MLODE workshop, and Steven Moran for his linguistic expertise.

References

[1] Chiarcos, C. (2010) Grounding an Ontology of Linguistic Annotations in the Data Category Registry. In: Proceedings of the 2010 International Conference on Language Resource and Evaluation (LREC)

[2] Chiarcos, C. (2013) OLiA - Ontologies of Linguistic Annotation. This issue.

[3] Chiarcos, C.; Nordhoff, S.; Hellmann, S. (eds.) (2012) Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata

[4] Comrie, B.; Key, M.R. (eds.) (2008) The Intercontinental Dictionary Series. Available online at `http://lingweb.eva.mpg.de/ids/` Accessed on 2013-06-10.

[5] Dryer, Matthew S.; Haspelmath, M. (eds.) (2011) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at `http://wals.info/` Accessed on 2013-06-10.

[6] Haspelmath, M. (2008) The typological database of the World Atlas of Language Structures. In: Martin Everaert and Simon Musgrave, editors, Typological databases. Mouton de Gruyter, Berlin.

[7] Ide, N.; Pustejovsky, J. (2010) What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, China.

[8] McCrae J.; Spohr D.; Cimiano P. (2011) Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In: The Semantic Web: Research and Applications, pp 245-259.

[9] Riechert, T. et al (2010) Knowledge Engineering for Historians on the Example of the Catalogus Professorum Lipsiensis. In: Proceedings of the 9th International Semantic Web Conference (ISQC 2010)

[10] Saulwick et al. (2005) Distributed tasking in ontology mediated integration of typological databases for linguistic research. 17th Conference on Advanced Information Systems Engineering (CAiSE0âĂŹ 05), Porto.

[11] Windhouwer, M; Schuurman, I; Wright, S.E. (2013) Collaboratively Defining Widely Accepted Linguistic Data Categories in the ISOcat Data Category Registry. ESWC2013 Workshop on Semantic Web Collaborative Spaces (SWCS2013), Montpellier.

[12] Relationship between ISO 639-3 and the other parts of ISO 639. In: ISO 639-3. SIL International `http://www.sil.org/iso639-3/relationship.asp` Accessed on 2013-06-10.

8. Appendix

```
PREFIX ids: <http://mlode.nlp2rdf.org/resource/ids/
    vocabulary/>
PREFIX wold: <http://wold.livingsources.org/>
PREFIX gold: <http://purl.org/linguistics/gold/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT *
WHERE {
        ?idsLang a ids:language.
        ?idsLang rdfs:label ?idsname.
        ?idsLang ids:hasIsoCode ?iso.
        ?woldLang a gold:Language.
        ?woldLang dcterms:title ?woldname.
        ?woldLang owl:sameAs ?silLink.
        FILTER(fn:substring-after(?silLink,’=’)=?
            iso)
}
```

Listing 1: SPARQL query to match IDS and WOLD languages

```
PREFIX ids: <http://mlode.nlp2rdf.org/resource/ids/
    vocabulary/>
PREFIX wals: <http://mlode.nlp2rdf.org/resource/
    wals/vocabulary/>
PREFIX gold: <http://purl.org/linguistics/gold/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT *
WHERE {
        ?idsLang a ids:language.
        ?idsLang ids:hasIsoCode ?idsIso.
        ?wals a wals:language.
        ?wals wals:hasIsoCode ?walsIso.
        FILTER(?idsIso=?walsIso)
}
```

Listing 2: SPARQL query to match IDS and WALS languages

```
PREFIX wals: <http://mlode.nlp2rdf.org/resource/
    wals/vocabulary/>
PREFIX wold: <http://wold.livingsources.org/>
PREFIX gold: <http://purl.org/linguistics/gold/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT *
WHERE {
        ?walsLang a wals:language.
        ?walsLang wals:hasIsoCode ?walsIso.
        ?walsLang dcterms:relation ?walsSil.
        ?woldLang a gold:Language.
        ?woldLang dcterms:title ?woldname.
        ?woldLang owl:sameAs ?woldSil.
        FILTER(?walsSil=?woldSil)
}
```

Listing 3: SPARQL query to match WOLD and WALS languages

```
PREFIX ids: <http://mlode.nlp2rdf.org/resource/ids/
    vocabulary/>
PREFIX wold: <http://wold.livingsources.org/>
PREFIX wals: <http://mlode.nlp2rdf.org/resource/
    wals/vocabulary/>
PREFIX gold: <http://purl.org/linguistics/gold/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT *
WHERE {
        ?idsLang a ids:language.
        ?idsLang ids:hasIsoCode ?idsIso.
        ?wold a gold:Language.
        ?wold owl:sameAs ?woldSil.
        ?wals a wals:language.
        ?wals wals:hasIsoCode ?walsIso.
        FILTER(fn:substring-after(?woldSil,’=’)=?
            idsIso && ?idsIso=?walsIso)
}
```

Listing 4: SPARQL query to match IDS, WOLD and WALS languages

```
PREFIX ids: <http://mlode.nlp2rdf.org/resource/ids/
    vocabulary/>
PREFIX wold: <http://wold.livingsources.org/>
PREFIX gold: <http://purl.org/linguistics/gold/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT *
WHERE {
        ?idsLang a ids:language.
        ?idsLang ids:hasIsoCode ?idsIso.
        ?idsEntry gold:inLanguage ?idsLang.
        ?idsEntry rdfs:label ?idsLabel .
        ?woldLang a gold:Language.
        ?woldLang owl:sameAs ?woldSil.
        ?woldWord gold:inLanguage ?woldLang.
        ?woldWord dcterms:title ?woldTitle .
        FILTER(fn:substring-after(?woldSil,’=’)=?
            idsIso)
        FILTER(regex(?idsLabel,?woldTitle))
}
```

Listing 5: SPARQL query to match words from WOLD with entries from IDS