WOLD, WALS and IDS

RDF conversion and interoperability of linguistic datasets of the MPI EVA Leipzig

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

Abstract. This paper describes the conversion into RDF, the internal structure, as well as the semantic content of three linguistic datasets of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthtopology. Two of the datasets where converted in the course of the MLODE 2012 workshop, while one is a pre-existent dataset converted by the MPI EVA. The description spans three datasets to illustrate similarities and differences, as well as common shortcommings in the conversion of linguistic datasets into RDF. Alongside the descriptions, the interoperability of the specific datasets and Linguistic Linked Open Data as a whole is examined and pitfalls common to the interaction of multiple datasets from different sources are discussed.

Keywords: multilingual, dictionary, language structure, linked open data, RDF, interoperability

1. Introduction

As an important body of linguistic research, the Department of Linguistics at the Max Planck Institute for Evolutionary Anthtopology[1] (MPI EVA) is working on a number of linguistic databases, different in scope, focus and internal structure. This paper serves to show best practices as well as pitfalls in stucturing and integrating linguistic datasets over different domains of interest via Semantic Web technologies. This will be done at the examples of 3 data sets of the MPI EVA published on the web: The World Atlas of Language Structures, a large-scale linguistic feature atlas; the Intercontinental Dictionary Series, a topically structured multilingual dictionary; and the World Loanword Database, providing loanword information for multilingual small scale vocabularies.

While describing the conversion and analyzation process, problems unique to Linguistic Linked Open Data (LLOD) will occur and tried to be solved, such as interlinking and interoperability of different linguistic resources, development of vocabularies and the final documentation of the data. Most of these are problems specific to the interaction of multiple datasets. The conversion of WALS and the IDS were done during the Multilingual Linked Open Data for Enterprises Workshop 2012[2] (MLODE) Code-a-thon. This happened project-wise, one dataset after the other. An approach like this has its drawbacks, which will be shown in the course of this paper. At the same time, it seems to be a common method, as the number of linguistic datasets in the Linked Open Data cloud is limited and one cannot link to data which is not available. Although the Linguistic Linked Open Data cloud[3] was heavily enlarged during the MLODE code-a-thon, this paper focusses on interoperability of newly converted and existing datasets, to ensure better quality conversion of linguistic datasets in the future.

This paper is structured as follows: The following three sections describe the structure and content of the three datasets. Section 5 will focus on the interoperability of the datasets, as well as interoperability of linguistic datasets as a whole. A special focus will be on interlinking and vocabulary design. Finally, the conclusion will be presented in section 6.

[1] `http://www.eva.mpg.de/lingua/`

[2] `http://sabre2012.infai.org/mlode`

[3] `http://linguistics.okfn.org/resources/llod/`

2. The World Atlas of Language Structures (WALS)

"The World Atlas of Language Structures (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."[1] It contains 192 structural *features* of languages, ordered in 144 different feature *chapters* such as number of genders, existence of different tenses and word order. Each feature contains a number of possible *values*. The feature "Order of Subject, Object and Verb", for example, has the values "SOV", "SVO", "VSO" etc. Tupels of the type `(language,feature)` are called *datapoints*, each containing the language specific value. WALS covers 2678 languages and most values exist for up to 500 datapoints. The data furthermore includes language names and alternative names, ISO639-3 codes, a classification into language families and genera and geographical coordinates.

The conversion of the dataset was based on a MySQL-database dump, but the data can also be retrieved online as CSV files. A vocabulary was developed to convert the main data, its datamodel shown in figure 1. URI formats and prefixes can be found in table 1. Although existing vocabularies like DCTERMS and WGS84 were used, most of the classes and properties were newly defined, due to the lack of a fitting existing ontology granular enough to describe the WALS data. Language resources are uniquely identified by a URI containing a three letter WALS code, an intern unique identifier. The ISO639-3 code was not used for this purpose, because the division of languages in WALS was made before the ISO639-3 standard was established and there is not always a 1:1 mapping between WALS code and ISO code[2]. Instead, existing ISO codes were used to link to a number of different language resources, like Lexvo[4], Glottolog/Langdoc[5] and SIL[6]. The features themselves are modeled as a property, connecting the language to a datapoint resource and containing the provenance of the feature information as a literal in a `dcterms:references` element. The values themselves are linked to the datapoints via the `wals:hasValue` property. The value information is contained in the `rdfs:label` and `dcterms:description` elements. For example, the value labeled "SOV" has the literal "Subject-object-verb (SOV)" in its `dcterms:description` element. The problem at this point is, that all the values contain this kind of information as literals. There is no well-defined structure, due to the heterogeneous nature of the different values. Thus, the information can be read and understood by humans, but is not easily parsed by machines. The discussed feature of word order for example, could be useful for automatical grammatical correction or inspection, but can not be genericly 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.

The resulting RDF version of WALS contains 499112 triples and over 6000 links to additional language resources. It was is published on `http://datahub.io/dataset/wals`. There was no Linked Data version or SPARQL endpoint available at the time of writing.

3. The Intercontinental Dictionary Series (IDS)

The Intercontinental Dictionary Series[7] is a multilingual dictionary organized in topical chapters to allow easy comparisons across languages. It contains 23 *chapters*, like "the physical world", "animals" or "food and drink" for 215 languages. These chapters order 1310 *entries*, essentially reference translations in English, French, Russian, Spanish and Portugese. Associated with these entries are the actual lexical items of the dictionary of which there are 280000. The IDS is a collaborative effort, compiled by its editors from a number of international sources.

For the RDF conversion, I had access to a PostgreSQL database dump of the IDS. Careful analyzation produced the data model shown in Figure 2 as a first step. Table 2 shows the URI formats and prefixes. To ensure interoperability, the tables containing the language data were most important and merged into the language class. It contains additional information like ISO639-3 codes and official (ISO) names of the languages, as well as alternative names and the source information of the data. This provenance information is especially important in the case of the IDS, because it is an international

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

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

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

[7] `http://lingweb.eva.mpg.de/ids/`

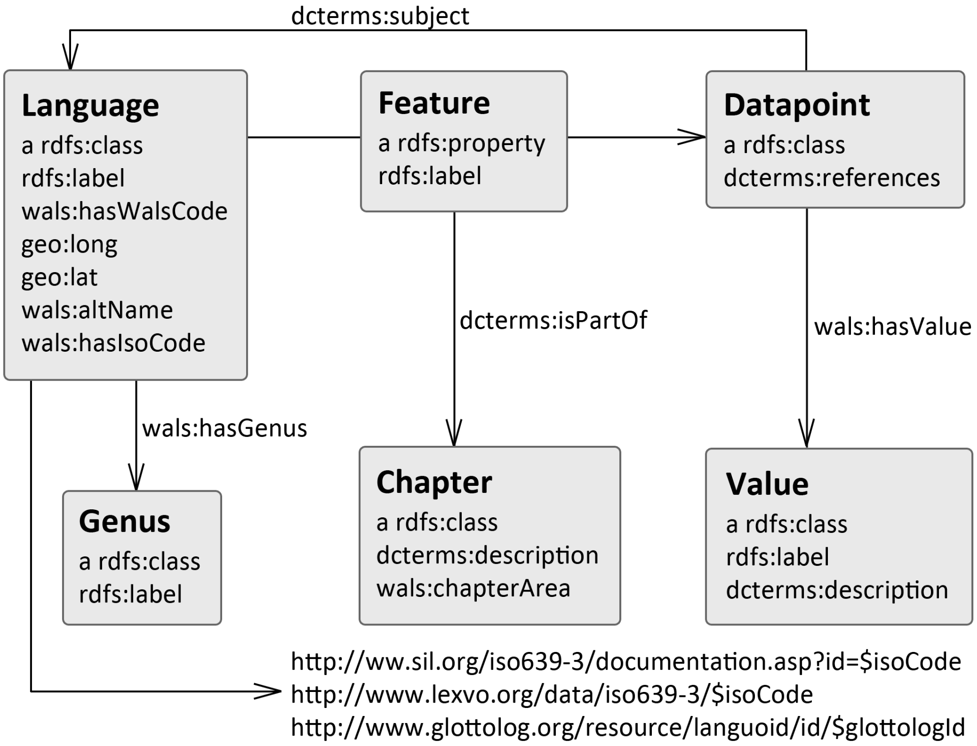

Fig. 1. WALS datamodel diagram

Table 1
WALS URI formats

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

collaborative effort. Data entry, compilation, consultation and sources are therefore converted to RDF as well, each as a literal in a 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 desireable for making granular queries. Language resources are ordered into classes, with the `inLangClass` property. Theses classes form a hierarchy of subclasses: 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 families or genera, like in the case of WALS. The entries themselves are devided into two kinds:

(a) The first kind of entries are the reference translations, containing the `ids:$translation` properties. Their label contains the english translation of the entry. They are furthermore linked to DBPedia Wiktionary[8]. Because of lacking part of speech information in the IDS, the linking ap-

[8] `http://dbpedia.wiktionary.org/resource/`

proach shown in[3] could not be used. The only check performed was for the existence of the Wiktionary resource. Because of this procedure, the correct word resource will be linked in many cases, but will be wrong in some of them.

(b) The second kind of entries are the lexical entries themselves. The `rdfs:label` of these resources contains the lexical entry. It is *not always normalized* and may contain multiple forms, seperated by semicolons, parenthesis or one or more minus signs. They also may contain further data like alternative forms or additional flections 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 entry are linked to the relevant chapter via `dcterms:isPartOf`.

To confirm to Linked Data principles, the language resources themselves 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 1984321 triples, with 216 links to each WALS, Glottolog and Lexvo, interlinking the language resources. There are additional 1832 links to DBPedia Wiktionary for the lexical entries and 22 links to WOLD semantic fields. The data was published at `http://datahub.io/dataset/ids`. There was no Linked Data version or SPARQL endpoint available at the time of writing.

4. The World Loanword Database (WOLD)

WOLD[9], is a database providing vocabularies of up to 2000 entries of 41 languages. Each entry features information about its loanword status, source words and associated meanings, thereby granting the possibility to find out donor languages and, through the means of geographical coordinates also provided, geographical distributions of these characteristics. Because donor languages are not necessarily part of the 41 languages analyzed, WOLD provides a varying amount of information for a total of 395 languages, some of them only mentioned by name. It is edited by Martin Haspelmath and Uri Tadmor and was published in 2009 under the Creative Commons Attribution 3.0 Licence[10]. An RDF+XML Version was added in 2010 as linked data by Robert Forkel, available on the WOLD website. There were 837828 triples and around 17000 links to DBPedia, WALS and SIL at the time of writing. There is no SPARQL-Endpoint on the project page available. An RDF dump was compiled and published on `http://datahub.io/dataset/wold`.

Although this dataset was not converted in the context of this paper, its structure will be described here for completeness and to highlight particular problems common to Linguistic Linked Open Data. Another reason is the relation of the project to the other datasets described in this paper, especially IDS, described in detail in section 5.3.

The RDF+XML version of the WOLD features 7 classes and uses existing ontologies, like GOLD[11] and the WordNet 2.0 schema[12], as well as SKOS[13]. Language resources only contain the name of the language and geographical coordinates as `kml:coordinates`[14]. ISO 639-3 can not be found explicitly in the data. They are contained in links of more widespread languages to SIL. Those links exist for 62% of the languages, limiting dataset interoperability on a language level. The same goes for links to the 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.

Language resources link vocabularies, which consist basically of a number of `dcterms:hasPart` properties linking word resources. Latter provide an orthographic representation, a lexical form and links to possible source words. The HTML representation again provides more information, like part of speech, comments and additional references, as well as a borrowed status, with assessments about the words likely source. Meanings are linked by words in a peculiar form: The `wn:sense` property links to a WordSense resource without an own identifier. Instead the URI is derived from a meaning resource and the identifier of the word attached as a fragment. So the `wn:sense` property of

[9] `wold.livingsources.org`

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

[11] `http://linguistics-ontology.org/gold`

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

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

[14] `http://www.opengis.net/kml/2.2`

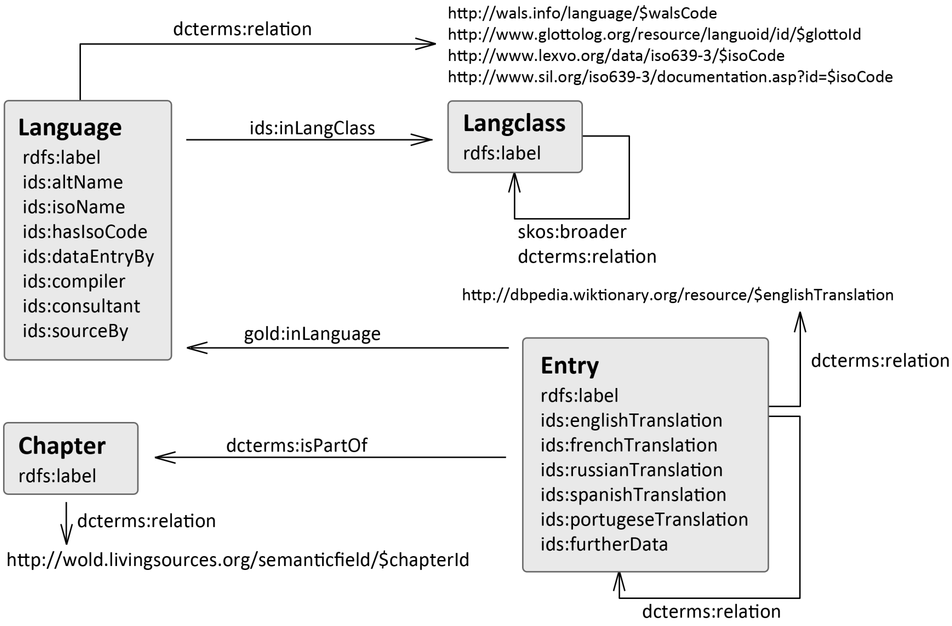

Fig. 2. IDS datamodel diagram

Table 2
IDS URI formats

Class	URI format
Language	`http://lingweb.eva.mpg.de/ids/language/$languageId`
Langclass	`http://lingweb.eva.mpg.de/ids/langclass/$langclassId`
Entry	`http://lingweb.eva.mpg.de/ids/entry/$entryId`
Chapter	`http://lingweb.eva.mpg.de/ids/chapter/$chapterId`
Prefix ids:	`http://lingweb.eva.mpg.de/ids/vocabulary/`

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 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 can not be accessed as Linked Data.

5. Interoperability

Interoperability is defined by [4] as devided into syntactic and semantic interoperability. RDF as a data format grants syntactic interoperability, as the standard is well-defined and can be processed by a number of existing tools. This is an important benefit, as established frameworks like Jena[15] allow easy and granular data access, aggregation and manipulation and web tools like OntoWiki[16] can be used to enhance collaborative research[5]. Semantic interoperability, as a means of "consistent interpretation of exchanged data"[4] on the other hand 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.

[15] `http://jena.apache.org/`
[16] `http://ontowiki.net/`

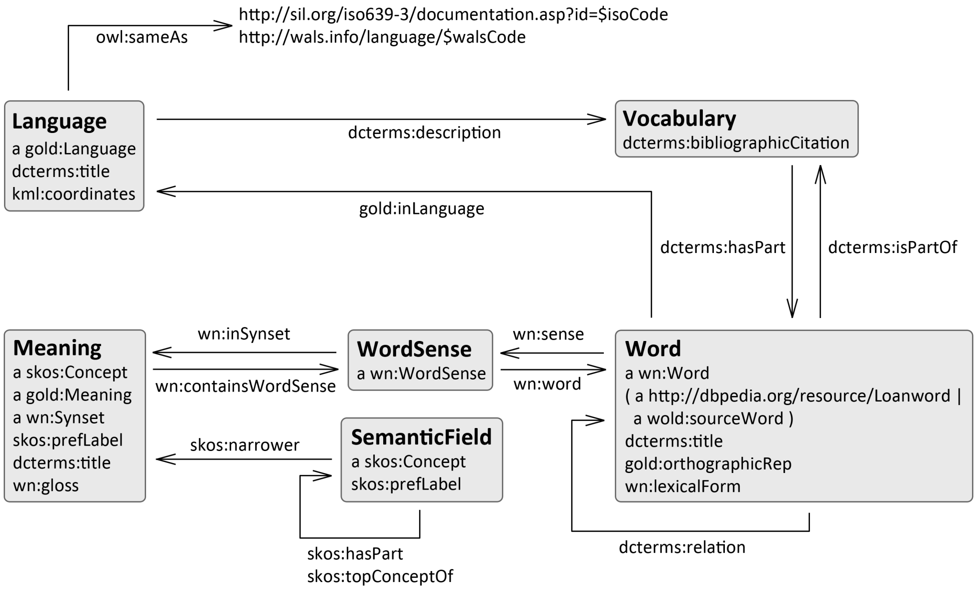

Fig. 3. WOLD datamodel diagram

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/`

Usual routines of vocabulary creation during dataset conversion must therefore be examined and adapted to the linguistic domain. Subsection 5.1 will show this in further detail.

Another hurdle of 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[6], 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 in specialized subdomains. Although efforts are made to elevate the issue by mapping different annotation schemes[7], the width of the linguistic domain makes this approach difficult and work-intensive. That is, if such a mapping is even feasible, considering that a *noun phrase* as a common feature may just be annotated differently over different datasets, but features like *consonant inventories*, as found in WALS, may not be mappable at all because of their specific focus.

(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, hindering interoperability. Furthermore, this definition must be clear to the compiler as well as the user of the data, which requires a 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 *data silos* to the public via the Semantic Web enhances interoperability and scope of language resources, but does not guarantee semantic interoperability. Therefore, even with structural interoperability given by RDF standards, without ontologies mighty enough to capture at least essential parts of linguistic subdomains, semantic interoperability will not be given.

5.1. Vocabularies

To grant automatic interpretation or semantically correct integration of different datasets, vocabularies and ontologies used must be compatible. At the time of writing, there is no commonly accepted and complete vocabulary for description of linguistic resources. The breadth of the linguistic domain may proof as 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, WN, SKOS, OWL and DCTERMS is encouraged and widespread, vocabularies are often defined bottom-up during the conversion into RDF. These definitions are furthermore made by Linked Data experts unfamiliar to the specific domain, like in the case of IDS in this paper and many other datasets converted for the MLODE 2012 workshop. The reasoning behind this practice is to avoid the definition of semantically inappropriate descriptors. This problem is typical to interdisciplinary use of Linked Data [5] and may be elevated in the future of LLOD. In the mean time, efforts like OLiA and ISOCAT 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 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 identificators, there would be no problem. The ISO 639-3 code was introduced for this specific purpose and has found wide acceptance.

It aims to provide language codes for every language, including living, extinct and ancient languages, as well as dialects[8]. The standard contains codes for over 7700 at the time of writing. While this number should be sufficient for most linguistic purposes, the problem still exists in data compilation. Linguistic field research may disagree about the specific boundary of small dialects, ancient languages or word forms of hard to specify origin. In Linked Data, this problem can be elevated by the use of specific properties. The `owl:sameAs` property there 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 language definition in different datasets. Rather, like suggested by the conversion of WALS and IDS in this paper, some kind of *relation* property, like `dcterms:relation` should be used.

Another important point is the question, what datasets should be linked. In the LLOD-Cloud, a number of datasets have been established as reference datasets for this purpose. These datasets are also linked by the datasets this paper describes, namely Glottolog/Langdoc, Lexvo and WALS. Lexvo can easily be linked, because its URIs contain the ISO 639-3 code. 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.

5.3. Practical interoperability on the example of the described datasets

To expand on the aforementioned points, interoperability was tested on the described datasets. The datasets were loaded into a triple store and then queried with SPARQL to achieve different goals. The queries can be found in the appendices. The first objective was matching of languages to integrate information about languages from the three different sources. This was done entirely by comparing ISO 639-3 codes, which proofed to be a hurdle, as mentioned in **??**. WOLD features ISO codes for only 62% of its language resources, severly limiting the possibilities of interlinking. They are further not directly contained as literals, but as parts of URIs, requiring string processing in the SPARQL-queries to extract them.

Table 4 shows the results for the language matching. 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 code. Percentages shown refer to the percentage of languages of the respective dataset matched to the languages of the other datasets. For WOLD, the percentage of languages featuring ISO codes was noted seperately, to

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)

reflect that the quantity of results is dependent on the existence of ISO codes.

Best results were obtained for matching WOLD and IDS languages to WALS. This results from the high number of languages contained in WALS. 78% of WOLD language resources with ISO codes and 83% of IDS language resources can therefore be enriched with WALS information, like 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 could be matched. The 41 languages WOLD focuses on were covered 100%. Integrating both datasets could yield interesting results. Because IDS is a multilingual dictionary and WOLD contains loan- and sourceword information, combining the data from both datasets is attractive. 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, this time to match on the level of individual words as well as language. One has to remember that IDS entries are not normalized and often contain mutliple lexical forms or other information. The query therefore used a `FILTER(regex(?idsEntry,?woldWord))` expression, which searches for occurences of the WOLD word string in the IDS entry label. This is not a precice, but in this case the only way to find a sufficient number of matches.

Although only 18% of WOLD languages could be linked to IDS languages, this query yielded 13030 triples. Out of a total number of 57926 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 Portugese. On the other hand, out of a total of 282671 IDS entries, 1% of them can be enhanced with loan- and sourceword information from WOLD. Although I am sure that these results can be improved by more exhaustive linking of language resources, there was 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 codes in all examined datasets. Again, all 41 languages focused by WOLD were among the results.

Although the datasets were converted for themselves and largely without interoperability as a specific aim, the results of the queries show interesting possibilities for data integration. If the Linked Data version of WOLD would contain all the information found in the HTML representation, integration with the IDS could be further enhanced or at least be done more precisely. The part of speech would be especially interesting for this feat. The points made in section **??** can not yet seen as proven, as I can not compare how many data could 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 focus of the projects, the greater coverage of WOLD languages with ISO codes shows the importance to incorporate commonly accepted codes into the datasets.

6. Conclusion

The datasets described in this paper present a useful contribution to the LLOD-Cloud. Data precision and 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 advantages as well as shortcomings in the datasets themselves as well as for Linguistic Linked Open Data as a whole. ISO 639-3 codes proofed to be essential for language interlinking and ensuring cross-dataset compatibility. The practical part of the paper has shown that sizeable 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. However, the advantages of RDF as a data format guaranteeing syntactic interoperability outweigh the disadvantages.

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.

References

[1] 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 2012-11-30.

[2] 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.

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

[4] 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.

[5] 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)

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

[7] 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)

[8] Relationship between ISO 639-3 and the other parts of ISO 639. In: ISO 639-3. SIL International, visited 01/01/2013 `http://www.sil.org/iso639-3/relationship.asp`

8. Appendices

```
PREFIX ids: <http://lingweb.eva.mpg.de/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://lingweb.eva.mpg.de/ids/
    vocabulary/>
PREFIX wals: <http://wals.info/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://wals.info/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://lingweb.eva.mpg.de/ids/
    vocabulary/>
PREFIX wold: <http://wold.livingsources.org/>
PREFIX wals: <http://wals.info/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://lingweb.eva.mpg.de/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