Managing Linguistic Resources by Enriching Their Metadata with Linked Data

Christina Hoppermann, Thorsten Trippel, and Claus Zinn

Dept. of Linguistics, University of Tübingen
firstname.lastname@uni-tuebingen.de

Abstract. The NaLiDa project aims at contributing to an infrastructure for the metadata-based description and access to linguistic resources and tools. When aggregating heterogenous metadata sets from various providers to provide a single and uniform point of access to the aggregation, data curation becomes a central issue. In this paper, we describe how we use authority files from the German National Library, available as Linked Data, to tackle this issue for metadata fields about persons, organisations, and subject classifications.

1 Introduction

In linguistics, there is an increasing amout of research data (*e.g.*, lexica, grammars, corpora, annotations, wordnets) and research tools (*e.g.*, part-of-speech taggers, parsers, annotation tools), complementing and creating a basis for published research results. While publications – when appearing in reputable journals or conference proceedings – obtain bibliographical metadata and enter bibliographic databases, there is no systematic procedure in place to manage and publish the underlying research data and tools. Recently, however, the attitude towards the sharing of research data is changing. The German Research Foundation, for instance, now requires research data to be kept for 10 years in a sustainable manner [2]: a resource needs to get captured on a lossless storage format and supplied with a description that allows easy identification and access.

Building a sustainable infrastructure for linguistic resources is a complex undertaking, with its major task to resolve various kinds of metadata issues. While there is an infrastructure evolving around the ISOcat Metadata Registry [4, 1], giving access to a community-based controlled terminology, the level of formalization for so-called *data categories* is targeted at human users. The style, hence, is rather informal as all definitions are given in natural languages such as English. Moreover, most data categories can take any string as value.

The infrastructure, while still in development, is already adopted by some data providers in linguistics. They distribute their metadata with schemas whose descriptors make all reference to ISOcat entries. This facilitates an automatic processing of metadata sets that adhere to such schemas. By building a faceted browser to give integrated access to aggregated metadata sets [5], we found a staggering amount of inconsistent entity naming. We found data curation necessary in the integrated collection, but also within single collections. In one instance, we identified three dozen lexical realisations of a single organisation, including the use of abbreviations, language variation (*e.g.*, English and French naming), and spelling errors. That made it impossible, for instance, to locate all linguistic resources originating from this institution with a single metadata-based or fulltext-based query. Similar problems exist for other contact metadata such as person and location names. This paper describes our use of Linked Data to address this issue.

2 The Authority Files of the German National Library

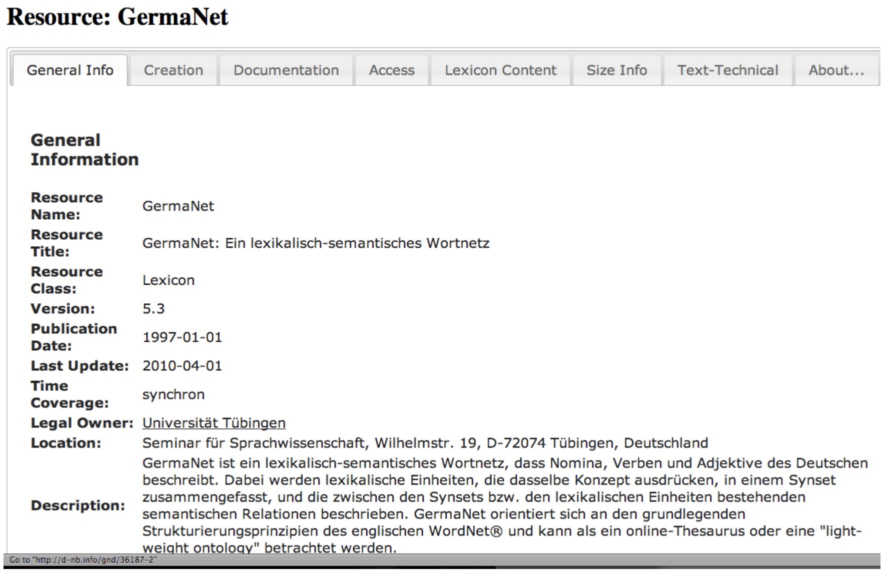
The German National Library (DNB) currently maintains three *authority files* for organisations, persons and subject headings, respectively. In the future, these will be merged into a common reference file (Gemeinsame Normdatei, GND). All data is available using the Linked Data Service of the German National Library, see `http://files.d-nb.de/pdf/linked_data_e.pdf`.

The *Corporate Body Authority File* (Gebiets- and Körperschaftsdatei, GKD) holds over 915.000 records of institutions. Each record gives an institution's preferred name, any known alternative names, a reference to its encompassing institution, and when available, to its predecessor and successor institution. Searching this database with the query "Tübingen", *e.g.*, lists over 2300 entries. Each of the three departments of the *Seminar für Sprachwissenschaft* are listed, together with a link to the Seminar. Many temporary organisations such as *Collaborative Research Centres* (Sonderforschungsbereiche) are also listed.

The *Personal Name Authority File* (Personennormdatei, PND) holds more than 3.6 million entries for persons, including the names of all persons that have authored a German publication or published their work in Germany. Around 1.8 million of them are *individualised entries*; these entries are associated with exactly one person and have attributes other than the person's name (with alternative names, if known): date of birth (and death), origin (with country code), profession, and references to entries in the SWD (see below). It is also possible to navigate back and forth between a PND entry and a person's publications. The PND holds nearly 4000 person entries with profession "linguist".

The *Subject Headings Authority File* (Schlagwortdatei, SWD) has 600.000 descriptors and about 700.000 (near-)synonyms to describe a collection item. About three quarters of the descriptors are proper names referring to persons, organisations, titles *etc.* There are about 115.000 hierachical and about 26.000 associative relations. In addition, all terms are classified in nearly 500 classes constituting 36 clusters. For linguistics, the classification lists around 40 different subfields, see `http://melvil.d-nb.de/swd/040742504`. There is also a mapping of SWD entries to the Dewey classification system, see `http://www.oclc.org/dewey/resources/summaries/default.htm#400`.

The sheer quantity of data in these three authority files, but also their high quality make them an ideal resource for referring to persons, organisations, and topics that occur in the context of metadata descriptions for research data.



General Information

Resource Name:	GermaNet
Resource Title:	GermaNet: Ein lexikalisch-semantisches Wortnetz
Resource Class:	Lexicon
Version:	5.3
Publication Date:	1997-01-01
Last Update:	2010-04-01
Time Coverage:	synchron
Legal Owner:	Universität Tübingen
Location:	Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen, Deutschland
Description:	GermaNet ist ein lexikalisch-semantisches Wortnetz, dass Nomina, Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische Einheiten, die dasselbe Konzept ausdrücken, in einem Synset zusammengefasst, und die zwischen den Synsets bzw. den lexikalischen Einheiten bestehenden semantischen Relationen beschrieben. GermaNet orientiert sich an den grundlegenden Strukturierungsprinzipien des englischen WordNet® und kann als ein online-Thesaurus oder eine "light-weight ontology" betrachtet werden.

Go to "http://d-nb.info/gnd/36187-2"

Fig. 1. Document View showing the metadata description of a linguistic resource.

3 Using Authority Files for Linguistic Resources

In the NaLiDa project, we harvest (via the OAI-PMH protocol, see [3]) metadata collections from a handful of linguistics departments. All data providers encode their metadata descriptions using CMDI (Component Metadata Infrastructure), an XML-based format where field descriptors need to refer to ISOcat data categories (see `www.clarin.eu/cmdi`). While we are now curating our own datasets wrt. person and organisation names, we cannot expect all institutions from which we harvest data to immediately curate their data. The standardisation of names, hence, has to take place at aggregation point.

The metadata that we harvest is stored – as is – into the document-based database CouchDB (see `www.couchdb.org`). Only when we build the indices for full-text and faceted search, we map all person and organisation names to the respective entries of the GKD and PND datasets. The table that maps the names given in the metadata files to the names of the two authority files was constructed manually (and is updated whenever necessary), using mainly the search and navigation devices of the German National Library. In the near future, we are going to mechanise this disambiguation process.

So far, we have not used the subject headings authority file for metadata descriptions as such data is surprisingly rare in the metadata sets we harvest.

When we display a metadata description for a linguistic resource, any organisation is now associated with a persistent URL linking to the organisation's entry in the DNB. Fig. 1 depicts the document view for the German Word-

Net resource. It shows that the value of the field “Legal Owner” (“Universität Tübingen”) is already linked to the respective GKD entry, while “Seminar für Sprachwissenschaft” in the field “Location” is not yet mapped to the respective entry. In the near future, we will also link city names (also appearing in contact metadata) to the respective linked data sets for geographic locations. Most references to languages and countries are already complemented with the respective ISO-639-x and ISO 3166 codes.

For the creation of new metadata, we are implementing an editor matching the design of the document view. The editor will use Linked Data to provide an auto-completion facility for each metadata field that asks for person, location, city, country, and language names. It will also also support auto-completion for other forms of controlled vocabulary as specified in the ISOcat data registry.

4 Discussion

Building a sustainable infrastructure for managing and accessing linguistic resources is hard. Proper metadata management is one of the central issues, and there are lessons to be learnt from the library sciences. Having authority files to keep track of person and organisation names supports the creation, management and access to linguistic resources. Research disciplines who aim at managing their research data should take these resources into account. In case the DNB authority files miss a person or organisation entry, local authority files should be created and maintained to keep track of such entries. So far, the SWD subject headings of the DNB are not used in the metadata sets we harvest. Here, we suspect that data providers are just unaware of this valuable resource. We equally advocate its use.

References

1. D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In *Proceedings of the 7th. International Conference on Language Resources and Evaluation (LREC)*, 2010.
2. Deutsche Forschungsgemeinschaft. Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission ”Selbstkontrolle in der Wissenschaft”. Denkschrift. Weinheim: Wiley-VCH. See `www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf`, 1998.
3. OAI-PMH. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol version 2.0 of 2002-06-14, Document Version 2008-12-07. `www.openarchives.org/OAI/2.0/openarchivesprotocol.htm`, 2008.
4. Int’l Organization of Standardization. Terminology and other language and content resources - specification of data categories and management of a data category registry for language resources (ISO-12620), Geneva. 2009. See `www.isocat.org.`
5. C. Zinn. Building a Faceted Browser in CouchDB Using Views on Views and Erlang Metaprogramming. In H. Kuchen, editor, *WFLP 2011 – Functional and (Constraint) Logic Programming*, volume 6816 of *LNCS*. Springer, 2011. See `www.sfs.uni-tuebingen.de/nalida/katalog/app/nalida/_design/nalida/index.html`.