Originally published in: McCrae, John P./Chiarcos, Christian/Montiel Ponsoda, Elena/Declerck, Thierry/Osenova, Petya/Hellmann, Sebastian (Eds.): Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LREC 2016 Workshop). 24 May 2016, Portorož, Slovenia. – Paris: European Language Resources Association (ELRA), 2016. Pp. 59-62.

Enhancing the Quality of Metadata by using Authority Control

Thorsten Trippel, Claus Zinn
Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
thorsten.trippel@uni-tuebingen.de, claus.zinn@uni-tuebingen.de

Abstract

The *Component MetaData Infrastructure (CMDI)* is the dominant framework for describing language resources according to ISO 24622 (ISO/TC 37/SC 4, 2015). Within the CLARIN world, CMDI has become a huge success. The Virtual Language Observatory (VLO) now holds over 800.000 resources, all described with CMDI-based metadata. With the metadata being harvested from about thirty centres, there is a considerable amount of heterogeneity in the data. In part, there is some use of controlled vocabularies to keep data heterogeneity in check, say when describing the type of a resource, or the country the resource is originating from. However, when CMDI data refers to the names of persons or organisations, strings are used in a rather uncontrolled manner. Here, the CMDI community can learn from libraries and archives who maintain standardised lists for all kinds of names. In this paper, we advocate the use of freely available authority files that support the unique identification of persons, organisations, and more. The systematic use of authority records enhances the quality of the metadata, hence improves the faceted browsing experience in the VLO, and also prepares the sharing of CMDI-based metadata with the data in library catalogues.

Keywords: Metadata quality, bibliographic metadata, authority records

1. Motivation

The Virtual Language Observatory (VLO) offers a faceted browser that helps users exploring linguistic resources at grand scale. At regular intervals, the VLO uses the OAI-PMH protocol to fetch metadata descriptions from about thirty partner organisations, ingests them into a single database, and offers a unified access to over 800.000 resources. While all partner organisations offer their metadata in CMDI, a huge data curation process is required to harmonise all data. Despite the common format, this data curation is by no means trivial. While some data providers make use of controlled vocabularies, for instance, to refer to country or language names, others use simple strings for this. Moreover, there are some data descriptors where strings are used by all parties, namely when referring to a person (say, as the creator of the resource) or an organisation (say, to describe where the resource has been created). Consider a user who uses the VLO to identify a resource in terms of the organisation it might be originating from. When the user asks the faceted browser to display a full list of organisations, the window in Fig. 1 (left) shows up. The user is confronted with, *e.g.*, four different spellings for the *Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)*. Whenever the user selects one of the four values, the metadata behind the other three spellings is automatically excluded from the search results, which is a rather unsatisfying user experience. As the screenshot indicates, this is not an isolated case. In the VLO, there are still hundreds of duplicates in the categories "organisation", "language" or "country", despite ongoing curation efforts such as CLAVAS [U3].

In the Library Sciences, where catalogues often contain millions of records from many different fields, the use of *authority files* is crucial. It allows librarians to associate alternative names with preferred ones. Fig. 1 (right) shows the authority record (taken from the *Gemeinsame Normdatei* of the German National Library, see below) for the BBAW. The record has a unique resource identifier, lists the organisation's preferred name, its alternative names, and other information such as the organisation's geographic location, or information about its predecessor and history. With an organisation being identifiable with a uniform resource identifier (URI) (such as `http://d-nb.info/gnd/2131094-4`), its name spelling becomes secondary. If all CMDI metadata providers would complement an organisation's name with a similar URI, the quality of all aggregated data would be greatly enhanced.

The unique identification of entity names is also highly relevant when linking CMDI-based metadata to external sources such as library catalogues and the linked open data initiative. It makes it possible to link the publications of a linguist with the linguistic resources he or she created.

2. Background

The German NaLiDa project[1] operates at the interface between subject field specific research infrastructures such as CLARIN and core infrastructures of research organisations such as libraries and computing centres. One aim is to define processes for ingesting metadata of linguistic resources (a subset of the VLO data) to the Tübingen Library Catalogue. In this process, we need to bridge the metadata standards used in the linguistics community with the dominant standards in the library world. The use of common URIs to refer to persons and organisations is such as bridge.

[1] NaLiDa is a project acronym abbreviating "**Na**chhaltigkeit **Li**nguistischer **Da**ten" (Sustainability of linguistic data).

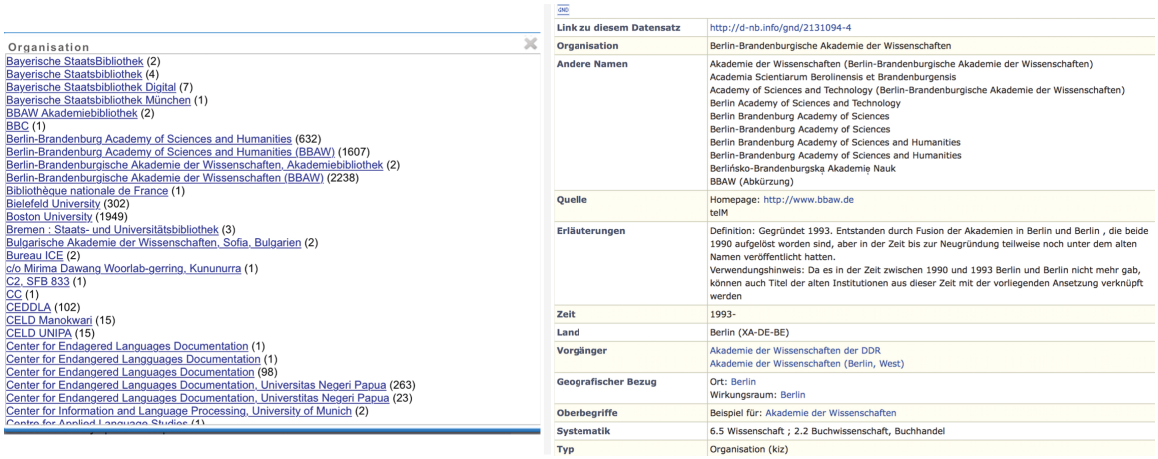

Figure 1: Organiation Duplicates in the VLO (left), the GND entry for the BBAW (right).

2.1. Metadata for language resources according to ISO 24622-1

The *Component MetaData Infrastructure (CMDI)* is a framework for the creation and use of metadata formats (CLARIN-D, 2012, page 19ff). Its abstract model follows an element-in-element, lego-brick approach to metadata modelling where schemas are defined by the selection and combination of predefined *data categories* and *components*. Data categories correspond to basic metadata elements or fields and are defined in the concept registry [U1], whereas components are hierarchically organized structures of data categories and components and are defined in the component registry [U2].

CMDI is the dominant framework for metadata in the CLARIN world, but it is also used by META-SHARE and other communities. At the time of writing, the CLARIN concept registry has about 1500 metadata terms and the CLARIN component registry offers over 1100 components with nearly 180 different public profiles (schemas).

With the rising numbers of resources described in CMDI, it is now time to adopt the use of authority files to uniquely describe the entities associated with the creation of the resources, and to hence complement string-based names for persons, corporate bodies *etc*. with authority records commonly used in the library world. This addresses the data heterogeneity issues described before.

2.2. The Use of Authority Files

In the library world, the use of authority files is good practise to identify persons, corporate bodies, but also subject headings. An authority file record gives a name in a standardised representation. It usually lists a person's (or organisation's) preferred name and complements it with alternative forms. Often, an authority record is associated with a unique resource identifier.

The **Integrated Authority File** (German: *Gemeinsame Normdatei (GND)*) is an international authority file for the organisation of persons, corporate bodies, conferences and events, geographic information, topics and works [U4]. It is maintained by the German National Library, and it has about 10 million entries, which includes over 2.5 million person names. The database is used widely in libraries, archives, and museums. It has a Creative Commons Zero (CC0) license.

The German National library also feeds the **Virtual International Authority File (VIAF)**, which is a joint project of several national libraries and is operated by the Online Computer Library Center (OCLC), see [U5]. The aim of VIAF is to link together the national authority files of all project members to a single virtual authority file. Each VIAF record is associated with a unique resource identifier and aggregates the information of the original authority records of the member states.

The **International Standard Name Identifier (ISNI)** is the "ISO certified global standard number for identifying the millions of contributors to creative works and those active in their distribution, including researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators, and more", see [U6]. It holds nearly 9 millions identities, including over 2.5 million names of researchers, and more than 500.000 organisation ids.

The US-American Library of Congress is another established authority file provider, see [U7]. More recent initiatives include the Open Researcher and Contributor ID (ORCID), see [U8], and ResearcherId, see [U9]. For geographical places, the GeoNames geographical database is widely used, see [U10].

All of these authority agencies attach a unique resource identifier to their records. Also, all agencies provide a RDF representation of records, so that it is possible to link together many data sources via the common format and the common use of identifiers. Note, for instance, that many Wikipedia biographical articles refer to the URIs of the aforementioned authority agencies.

3. Adding Authority Information to CMDI-based metadata descriptions

CMDI is a flexible framework making it easy to add provisions for authority records. For this, it is nec-

(a) GND record: Thorsten Trippel

(b) GND record: Claus Zinn

Figure 2: The authors' names in the GND.

essary to use data descriptors in the CLARIN concept registry: `/issuingAuthority/` (added to the concept registry) and `/id/`. Values for the concept `/issuingAuthority/` must stem from a controlled vocabulary referring to the authority institutions that we currently support. Currently, we include VIAF, GND, ISNI, ORCID, LC and `geonames.org`.

In the CLARIN component registry, we define the component `/AuthoritativeID/`, which holds the aforementioned two data descriptors, and `/AuthoritativeIDs/`, which brackets one or more occurrences of `/AuthoritativeID/`.[2] References to authority files are modelled as pairs of unique resource identifier and authority, where we use the controlled name of the authority registering the identifier. Modelling the authority reference as a pair of identifier and issuing institution makes it easy to add other authorities when required.

Fig. 3 depicts the use of the new descriptional means when referring to a person. In the given case, we associated three different authority records to the string denoting the person *Erhard Hinrichs*. It shows that about 60% of all names occurring in our local CMDI instances can be complemented with information from authority records stemming from GND, VIAF or ISNI. Notably, all researchers with a PhD are covered. Note that all organisations in our local CMDI instances have corresponding GND, VIAF, or ISNI records.

In sum, the curation effort is manageable. Having modified the CMDI profiles, the metadata instances must be adjusted to adhere to their new profiles. First, all instances must now have a reference to the modified profile. Second, when persons, organisations and locations are given (usually as strings), those are complemented with corresponding references to their respective authority records. Given there is a (hand-made) table associating name strings with authority records, an XSLT style-sheet can be written to mechanise the updating of the CMDI instances.

Having associated authority file information with person names (*i.e.*, strings), some other bits of information often included in the CMDI metadata may become redundant, but not necessarily so. In fact, the affiliation of a person given in the original CMDI metadata may well be different to the affiliation of this person given in the authority record. The first affiliation indicates where the person worked when the resource in question was created; this is often more relevant than the person's affiliation given in the authority record.

[2]Due to the recursive nature of profiles, about a dozen of other CMDI components that contain references to names (such as `/Contact/` or `/Funder/`) were modified to include the new component. Note that, at the time of writing, the new concepts and components reside in the private space of the CLARIN registries.

4. Discussion

The use of authority files greatly improves the quality of the CMDI-based metadata. Persons and corporate bodies are now uniquely identifiable. When CMDI data providers adopt authority files, we will see two main benefits: (i) an improvement in search through aggregated data sources within the CLARIN Virtual Language Observatory (especially wrt. organisations), and (ii) a better linking to library catalogues which use the same authority file information. The latter makes it possible to find a researcher's entire work (traditional publications and research data) with a single query. Data sharing at the URI level pays off.

We have seen that authority records may contain information about a person's birthdate, sex, academic degree, or profession. The record may give a reference to a geographical location (where the person works or has worked). Some of this information will not be up to date, see for instance,

```
<Person>
    <firstName>Erhard</firstName>
    <lastName>Hinrichs</lastName>
    <Role>Projektleiter</Role>
    <AuthoritativeIDs>
        <AuthoritativeID>
            <id>http://viaf.org/viaf/37069402</id>
            <issuingAuthority>VIAF</issuingAuthority>
        </AuthoritativeID>
        <AuthoritativeID>
            <id>http://d-nb.info/gnd/143840657</id>
            <issuingAuthority>GND</issuingAuthority>
        </AuthoritativeID>
        <AuthoritativeID>
            <id>http://isni.org/0000000118749683</id>
            <issuingAuthority>ISNI</issuingAuthority>
        </AuthoritativeID>
    </AuthoritativeIDs>
</Person>
```

Figure 3: Example fragment for the new encoding for a person name.

the informaton given in Fig. 2(a).[3] Here, exitisting CMDI metadata may well overwrite or complement the information associated with a person's authority record.
With over 2.5 million person names in the GND, there are entries that share the same name. By coincidence, a GND search for "Claus Zinn" shows two entries. The second entry is less specific than the one given in Fig. 2(b); it may well be an unwanted duplicate. In fact, associating the correct authority file with a given name is often facilitated by the additional information the record contains, in particular, the person's profession or associated publications.
To our knowledge, only libraries can directly enter or update existing authority file information. Note that the German National Library allows users to easily request a correction or actualisation of their GND entries (each GND record is displayed with an action "request correction").

So far, the CMDI community makes little use of metadata standards and controlled vocabularies used elsewhere. There are three major avenues to develop CMDI toward other metadata standards, and to bring CMDI closer to the library world, and subsequently toward the Semantic Web:

- making available tools that map CMDI to other metadata standards, in particular, towards the dominant standards in the library world such as Dublin Core [U11] and MARC 21 [U12].
- making available conversion tools that convert the XML-based CMDI representations to RDF-based representations where all information is expressed in terms of RDF triples.
- using unique resource identifiers to refer to persons, corporations, and geographical places.

Existing work tackles the first and second aspect of opening up CMDI to the metadata world. (Ďurčo and Windhouwer, 2014) propose a conversion from CMDI to RDF, and (Zinn et al., 2016) propose crosswalks between CMDI-based profiles and the library metadata standards MARC 21 and Dublin Core. In isolation, none of the work yields results that a librarian will be entirely happy with. Having a CMDI-based record converted to MARC 21 helps its ingestion in the library catalogue, but without authority information the new information is not linked to any prior information in the catalogue (*e.g.*, common author or common publisher). Similarly, while having a CMDI-based record be expressed in RDF has a number of advantages (*e.g.*, common data format with other data sources, RDF-based technology for storing or querying data sets), a true conversion of CMDI-based RDF data requires data sharing at the URI level.
In this paper, we have addressed the third aspect, incorporating authority records into CMDI-based metadata descriptions. This vastly improves the conversion to MARC 21 and RDF, and it also strengthens the links to other datasets. We encourage other CMDI metadata providers to follow our steps.

[3]The GND record has Trippel's geographic location given as *Bielefeld*, which was true at a time when he wrote his PhD thesis.

We also encourage the CLAVAS initiative, which seeks to produce a curated list of organisations based on the CLARIN VLO, see [U3], to associate with each organisation a reference to the respective record from the GND database. In fact, the many alternative names present in an organisation's GND record could be used to partially automate the mapping process.

For the long term, the CLARIN consortium may want to consider a metadata policy that propagates (or even enforces) the use of authority records in CMDI-based metadata.

Note. All our CMDI metadata, enriched with authority file information, will soon be available in the CLARIN VLO.

Acknowledgments. The NaLiDa project has been funded by the German Research Foundation, reference numbers DO 1346/4-2, WA 3085/1-2, and HI 495/4-2.
We would like to thank the anonymous referees for their comments, which helped improve this paper considerably.

Web Resources

[U1] The CLARIN Concept Registry, see `openskos.meertens.knaw.nl/ccr/browser`
[U2] The CLARIN Component Registry, see `catalog.clarin.eu/ds/ComponentRegistry`
[U3] The CLAVAS OpenSKOS Vocabulary Service, see `openskos.meertens.knaw.nl/clavas`
[U4] The Integrated Authority File of the German National Library, see `www.dnb.de/EN/Standardisierung/GND/gnd_node.html`
[U5] The Virtual International Authority File, see `viaf.org`.
[U6] The International Standard Name Identifier, see `isni.org`.
[U7] The Library of Congress Control Number, see `id.loc.gov/authorities/names.html`.
[U8] The Open Researcher and Contributor ID, see `orcid.org`.
[U9] ResearcherId, see `www.researcherid.com`.
[U10] The GeoNames database, see `geonames.org`.
[U11] The Dublin Core Metadata Initiative, see `www.dublincore.org`.
[U12] The MARC 21 standard, see `www.loc.gov/marc/bibliographic`.

5. Bibliographical References

CLARIN-D. (2012). The CLARIN-D user guide. `http://media.dwds.de/clarin/userguide/text`.
ISO/TC 37/SC 4, (2015). *ISO 24622-1:2015(en) Language resource management Component Metadata Infrastructure (CMDI) Part 1: The Component Metadata Model*. International Organization for Standardization.
Ďurčo, M. and Windhouwer, M. (2014). From CLARIN component metadata to linked open data. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 24–28. Co-located with LREC 2014. 26.-31 May 2014, Reykjavik. ELRA.
Zinn, C., Trippel, T., Kaminski, S., and Dima., E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In *Proceedings of LREC 2016, Portorož*. ELRA.