Crosswalking from CMDI to Dublin Core and MARC 21

Claus Zinn, Thorsten Trippel, Steve Kaminski, Emanuel Dima
Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
firstName.lastName@uni-tuebingen.de

Abstract

The Component MetaData Infrastructure (CMDI) is a framework for the creation and usage of metadata formats to describe all kinds of resources in the CLARIN world. To better connect to the library world, and to allow librarians to enter metadata for linguistic resources into their catalogues, a *crosswalk* from CMDI-based formats to bibliographic standards is required. The general and rather fluid nature of CMDI, however, makes it hard to map arbitrary CMDI schemas to metadata standards such as Dublin Core (DC) or MARC 21, which have a mature, well-defined and fixed set of field descriptors. In this paper, we address the issue and propose crosswalks between CMDI-based profiles originating from the NaLiDa project and DC and MARC 21, respectively.

Keywords: Metadata formats, crosswalks, bibliographic metadata

1. Motivation

The NaliDa project aims at constructing an infrastructure for the long-term archival of linguistic resources with technology and workflows that are manageable and sustainable. The infrastructure relies on resource providers to use rich but standardized descriptional means to associate their resources with metadata; it entrusts the university's central computing services with the long-term archiving of linguistic resources; and it asks the university library to ingest all resources' metadata into their catalogues so that interested parties can easily search for them. The project aims at fostering a close cooperation between library and computing services where library catalogues are connected with the research data repositories of the computing centre, so that users can profit from easy-to-use access points. With the new infrastructure in place, research data will move from the individual research departments to the infrastructure institutions of the university, which have the luxury of longer time spans and more stable financial, personnel and computing resources.

Making linguistic resources accessible to researchers requires their accurate description with metadata. In our workflows, resources are not described by librarians but by the resource creators (who know their resources best). They make use of the Component Metadata Infrastructure (CMDI) to describe their resources with rich metadata using a significant amount of linguistic vocabulary. In turn, this allows linguists to better search for and identify linguistic resources of their interest. In fact, for each type of resource that is being created in our department, we have devised a purpose-built CMDI profile to allow users to describe resources of this type is the best possible way.

Making accessible linguistic resources via standard library catalogues is not trivial, because their CMDI-based metadata records cannot be simply ingested into existing library catalogues. In the library world, there are only a few established metadata standards that library cataloging depends on. For librarians, CMDI is not an acceptable format as it does not fit their cataloging infrastructure. It is hence necessary to convert *discipline-specific* and *rich* metadata such as CMDI-based formats to library-specific formats such as MARC 21 or Dublin Core (DC). Those formats serve all disciplines, and therefore, their metadata descriptors are more abstract or universal than their CMDI-based counterparts.

The conversion from CMDI-based to bibliographic formats is challenging; it must address the risk of loosing valuable metadata information when the target formats are ignorant of discipline-specific terminology and needs.

The remainder of the paper is organised as follows: Sect. 2. gives background information about the metadata framework used in CLARIN, the metadata standards MARC 21 and Dublin Core used in the library world, and the crosswalk methodology. Sect. 3. then presents the difficulties when mapping between metadata standards, and the solution paths we have taken to address the issues. Sect. 4. discusses our work, and Sect. 5. concludes. The Appendix shows an example representation of a linguistic resource in the bibliographic formats Dublin Core and MARC 21.

2. Background

2.1. The Component Metadata Infrastructure

CMDI provides a framework for the creation and use of self-defined metadata formats (CLARIN-D, 2012, page 19ff). Its abstract model follows a lego-brick approach to metadata modeling: given the requirements of metadata modelers for a domain model, a schema is defined by the selection and combination of given, predefined *data categories* and *components*. Data categories correspond to basic metadata elements or fields, whereas components are a hierarchically organized structure of data categories or components. When data modelers cannot find data categories or components that fit their purpose, they can define their own making use of the CLARIN concept [U1] or component registry [U2], also see (Broeder et al., 2010).

In summary, CMDI-based metadata has the following properties: (i) each data category and component has a unique, persistent identifier; (ii) each data category has a semantic definition, and components are defined in terms of their parts; (iii) there are constraints imposed on the value a data descriptor can take, *e.g.*, free text, numeric range, date, or a controlled vocabulary; (iv) a CMDI schema specifies for

each element whether it is mandatory, optional, or whether it may occur multiple times; and (v) a CMDI schema organizes the metadata elements in hierarchical parent-child relationships.
In contrast to many other standards, CMDI is a grassroots movement. Data categories and components are created by many volunteers from different sub-disciplines, with varying expertise in metadata policies and issues. And there are little power structures in place that strive for a high-quality and uniform standard for the definition of CMDI elements. The CLARIN registries testify the uncontrolled growth of CMDI elements, making it hard if not impossible to compare arbitrary CMDI-based schema to other metadata standards. Another concern is the *usage* of CMDI concepts and components. Usually, CMDI profiles do not come with extensive documentation or guidebooks that help metadata providers to use the descriptors in the intended way, and often, the semantic definition of a data category is vague or incomplete. In fact, we have seen CMDI-based descriptions where vital information is missing or where it is given in the wrong metadata field.

2.2. Bibliographic metadata standards

The systematic description of books, sound and video recordings but also other artifacts has a long tradition in the Library Sciences. For the proper cataloging of such items, information about the resources' creators, titles, formats, dimensions, subject terms *etc.* are being aggregated in bibliographic records. There exists a number of dominant metadata standards to record such information such as MARC 21 and Dublin Core.

The Dublin Core (DC) metadata standard, initiated by the Dublin Core Metadata Initiative (DCMI) in the 1990s, provides a small *set* of terms to describe any kind of web and physical resources [U3]. DC is a lightweight standard with only 15 core elements, all of which are optional and repeatable to use [U4]. DC is often considered minimal common ground for metadata exchange between different communities of practice. Qualified Dublin Core adds *qualifiers* to the core elements of the standard [U5]; they help making the meaning of an element more specific, given a community of practise, or help denoting a scheme to aid the interpretation of an element value.

MARC 21 (**MA**chine **R**eadable **C**ataloging) is a rich and widely used proprietary encoding standard developed specifically for the description of bibliographic resources and for the facilitation of the exchange of bibliographic information among libraries [U6]. The history of the MARC standard dates back to the 1960s. It became the national standard of bibliographic data in the United States in 1971, and by today, it is the predominant standard for library cataloging worldwide, offering an unmatched level of granularity. Like Dublin Core, MARC 21 constitutes a non-hierarchical system. A MARC record typically consists of three sections: a fixed-length leader, a directory, and a set of data fields, which hold the descriptive metadata for the resource. A data field is identified by a *three-character tag*, *indicators* that supplement the data found in the data field, and a number of subfields (examples are given below).

There are other standards such as *Metadata Object Description Schema (MODS)*, which was designed by the Library of Congress to be less complex than the MARC format, but more expressive than Dublin Core, see [U7]. There is also *Metadata Encoding and Transmission Standard (METS)*, a metadata standard devised to describe the resources held in digital libraries, see [U8]. Archives often use the *Encoded Archival Description (EAD)* to describe their holdings, see [U9]. In contrast to bibliographical formats, EAD is a hierarchical metadata format designed to describe a given collection as a whole but also to give a detailed multi-level inventory of the collection.

2.3. Crosswalking

The sharing of information about resources across institutions that adopt different metadata standards requires a *crosswalk*. A crosswalk attempts to map the elements in one metadata format or schema to the semantically equivalent elements in another schema. In the Library Sciences, there are a number of established metadata crosswalks, for instance, from DC to MARC 21 [U10], from MARC 21 to DC [U11], from DC to EAD [U12], and from MARC 21 to EAD [U13].
Mapping one metadata standard to another is not an easy task (Pierre and LaPlant, 1998; Godby et al., 2004; Woodley, 2008). Often, there is no one-to-one mapping from a given data descriptor of the source metadata format to another descriptor in the target format. Often, no such mapping can be found, or many-to-one or one-to-many mappings need to be constructed.

In the library world, CMDI is practically unknown. Making accessible linguistic resources via library catalogues requires a format understood by librarians and their catalogue software. We therefore need to convert CMDI-based metadata to a bibliographic metadata standard. For this, we need a crosswalk that tables the relationships and equivalences between the metadata fields of the two standards.

3. CMDI Crosswalks to DC and MARC 21

The crosswalks will be unidirectional from CMDI-based profiles to DC and MARC 21.

3.1. Obstacles

The (grass roots) nature of CMDI-based profiles makes it hard if not impossible to define a general mapping from arbitrary CMDI-based schemas to DC and MARC 21. The difficulties are manifold.
While MARC 21 and DC have a mature, well-defined and relatively stable set of descriptors, CMDI schemas may refer to data categories that have a vague definition, or to elements that are rarely used by other schemas. In fact, metadata modelers often define their own descriptional means in the CLARIN registries rather than using predefined ones. At the time of writing, the CMDI framework has about 1500 metadata terms in the CLARIN concept registry, and over 1000 components and about 180 schemas in the CLARIN component registry. In comparison, DC has only 15 categories, and the fine-grained MARC 21 standard

has about 200 data fields, each of which can have many subfields. As a consequence, the crosswalks between CMDI-based schemas and the bibliographic standards will often involve many-to-one and many-to-none mappings.

The different abstract models are another source of problems. CMDI follows an *element-in-element model* so that the meaning of a data descriptor is, at least partially, derived from its hierarchical context. Consider the data category `/description/`, whose context embedding determines, *e.g.*, whether it is used to describe the resource as such, or the organization that created the resource, or the technical access to the resource, or the format of the resource. DC and MARC 21 both follow a *flat* model; here the metadata record is a simple *set* of property-value pairs. The single element `dc:description`, *e.g.*, is used to describe the resource as such, and not any particular aspect of it. In MARC 21, there are different fields available to cover those aspects covered by the contextual CMDI use of `/description/`. To address these issues, our mapping will be confined to NaLiDa-based profiles where we are aware of all contextually relevant information.

3.2. Refactoring of CMDI-based NaLiDa profiles

At the start of the metadata conversion project, the NaLiDa project had around 20 different NaLiDa profiles to describe a wide variety of linguistic resources: corpora, dictionaries, experiments, frequency lists, language documentation, lexica, named entity lists, speech corpora, thesauri, treebanks, web services *etc.* All profiles had a large number of components (and hence data categories) in common, such as general information about the resource, information on the resource' access, its creation, the project it is originating from, or project-related publications. Naturally, there were also elements to describe very specific aspects of, say, a dictionary, a frequency list, or a lexicon.

During the conversion project, it was decided to radically cut down the number of NaLiDa profiles to four main schemas to describe four main classes of resources: lexical resources, text corpora, experiments, and tools. Also, the remaining four profiles now share a larger resource-independent part, and each have an individual resource-specific part. With this refactoring of NaLiDa-based CMDI profiles, the mapping to Dublin Core and MARC 21 became more manageable.

Fig. 1 shows the main structure of the CMDI profile for lexical resources [U14] (ignoring the arrows indicating the mapping). The profile's main subtrees are labeled `/GeneralInfo/`, `/Project/`, `/Access/`, `/Creation/`, `/TechnicalInfo/`, `/LexicalResourceContext/`, `/Documentations/`, and `/Publications/`. Note that the latter two subtrees are not given as no mapping to DC is possible; also note that the other three NaLiDa-based CMDI profiles share all subtrees, except the resource-specific one, namely, `/LexicalResourceContext/`.

3.3. The CMDI to DC crosswalk

By taking into account the arrow information, Fig. 1 shows a crosswalk between a CMDI-based schema for lexical resources to DC. Given that DC has only 15 elements, a considerable amount of information is lost during the mapping. Many CMDI descriptors have no equivalent DC element, for instance, information about the version, location or modality of the resource, or project-related information about the project's funders or cooperation partners.

The mapping also shows many-to-one mapping where the data element `/Description/` occurs in different contexts (subtrees), but looses this context when uniformly mapped to `dc:description`. The crosswalk in Fig. 1 also shows that there are cases where the two CMDI fields `/firstName/` and `/lastName/` must be mapped to a single DC field `dc:creator` (or `dc:publisher`). Moreover, there are a number of CMDI elements such as `/DistributionMedium/` and `/TotalSize/` that carry information mappable to `dc:format`.

3.4. The CMDI to MARC 21 crosswalk

As we have noted, there is an established, bidirectional crosswalk between DC and MARC 21. Going from CMDI to MARC 21 via DC, however, propagates the information loss suffered from the previous mapping. Given the relative richness of MARC 21 with respect to DC, we propagate a direct mapping from CMDI to MARC 21. We highlight the mapping for the main data descriptors; for the following discussion, please consult Fig. 3 in the Appendix.

The MARC field with tag **100** associates the main personal name with the resource; the subfield **$e** can be used to further describe the author's role (author, funder, sponsor, illustrator, corrector), and the subfield **$u** can hold the person's affiliation. For our NaLiDa profiles, we assign the first creator to tag **100**, all other creators are stored in the MARC field **720**. The MARC field **245** holds the resource' title information. It may include a subtitle or the resource's medium (*e.g.*, "sound recording", or "electronic resource"). The MARC field **256** can be used to store computer file characteristics. We use this field to hold all information of the CMDI component `/SizeInfo/`. The MARC field **260** holds information "relating to the publication, printing, distribution, issue, release, or production of a work". The MARC field **500** is used for general notes to describe the nature, form, or scope of the item, and serves as a *fallback field*, where we enter all information for which no other MARC fields exist. The MARC field **505** holds "titles of separate works or parts of an item or the table of contents". We use this field to store information from the CMDI component `/DeploymentToolInfo/`. The MARC field **506** holds restrictions on accessing a resource. The subfield **$a** holds legal, physical, or procedural restrictions imposed on individuals wishing to see the described materials; the subfield **$u** holds the URL where the resource can be e-accessed, and the subfield **$f** holds standardized terminology for describing any access restrictions. This field is used to store most of the information from the CMDI component `/Access/`. The MARC field **520** holds unformatted information that describes the scope and general contents of the materials. Here, we map all values from the CMDI Component `/Description/`, taken from the `/GeneralInfo/` context, onto this field. The MARC field **536** holds information about funding. Here, we store the values of the CMDI element `/Funder/`. The

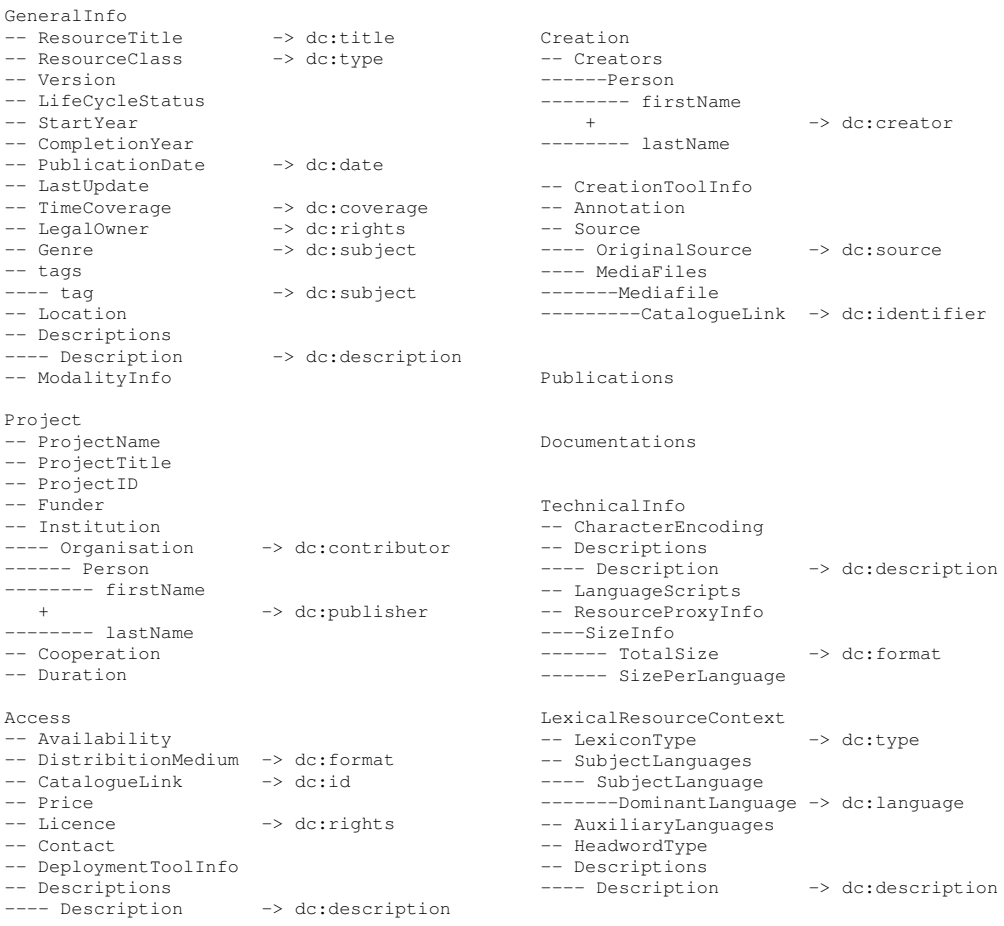

Figure 1: Mapping the leafs of the LexicalResourceProfile to DC.

MARC field **546** holds "textual information on the language or notation system used to convey the content of the described materials". In subfield **$a**, we store information from the CMDI component `/SubjectLanguages/`; in subfield **$b**, we map CMDI-based information from `TextTechnical.LanguageScripts`. The MARC field **653** holds uncontrolled index terms. We map to this field values from the CMDI field `/Genre/` as well as values stored in the CMDI field `/tag/`. The MARC field **856** holds information about electronic location and access. We use this field to hold information about the resource's mimetype (subfield **$q**) and URL (subfield **$u**).

4. Discussion

Incompatible metadata descriptions hinder effective search (Godby et al., 2004). To facilitate search across institutional or disciplinary boundaries, it is necessary to define crosswalks between the various metadata standards. In the CLARIN context, there are a number of crosswalks in use. When making metadata available through OAI-PMH harvesting, CLARIN repository providers must complement CMDI-based descriptions with metadata in DC. Given the nature of the CMDI framework, there is no universal crosswalk available that maps arbitrary CMDI metadata to DC. In line with our work, the individual CMDI profiles, and potentially, contextual information, must be taken into account to achieve a high-quality mapping to DC.[1]

In (Kemps-Snijders et al., 2012), the authors describe the use of CMDI at the Meertens institute. With an existing CMDI-based infrastructure in place, in particular, the Virtual Language Observatory (VLO) as central hub to linguistic resources, the Meertens institute converted metadata from various formats into CMDI. Following their account, which unfortunately gives no information about the source metadata formats, a number of custom scripts were developed for a bulk conversion process to CMDI. As a result, metadata for around 250.000 songs of the *Liederenbank* were transformed into a CMDI-based format, manually controlled for quality, and subsequently ingested into the Meertens repository. Through OAI-PMH harvesting all records are now available within the CLARIN VLO.

Given the expressiveness of CMDI, the conversion to less expressive metadata formats seems like a step back, unless it is either required (as in the OAI-PMH case) or unless

[1]On the CLARIN site `https://www.clarin.eu/content/oai-pmh-cmdi`, we find: "[...] how should I map my CMDI descriptions to the dublin core format that is compulsory when using OAI-PMH? The answer is: you probably know this the best, there is no single answer to this. It's probably a good idea to use common sense". Alternatively, it is proposed to create a minimalistic DC description that only consists of a single field, `dc:identifier`.

it has other benefits. The NaLiDa project, which funded this work, acts at the interface between the Linguistics Department and the infrastructure institutions of the University of Tübingen, namely, its library and its computing centre. From the linguistic perspective, the possibility to have metadata about our linguistic resources accessible and findable using a standard library catalogue search is attractive. It allows researchers unaware of the VLO to discover resources, sometimes by a happy coincidence. A library search with some author query will not only yield all the traditional publications (books, book chapters, articles *etc.*) associated with the given author, but now also returns linguistic resources that are associated with this name.

By mapping CMDI to other metadata formats, we have learned some lessons about metadata, which helped improve our existing policies and schemas. One piece of information that is now an integral part of our CMDI profiles is the use of authority files. In (Trippel and Zinn, 2016), we propagate the use of authority records, which have a long tradition in the Library Sciences, to identify persons, institutions and geographical places at the level of unique resource identifiers. Such as policy, if adopted in the CLARIN world, would help improve the (faceted) search experience in the Virtual Language Observatory.

5. Conclusion

The framework character of CMDI prevents us from defining a general mapping of CMDI-based metadata to bibliographic standards such as DC and MARC 21. For our NaLiDa data, we attempted to minimize the information loss when mapping to DC. All 15 elements were used, and arguably, sometimes overused (e.g., all kinds of descriptions were mapped to `dc:description`). Given the richness of MARC 21, we encountered only little information loss. Information that could not be stored in any other fields is stored in the data field **500**.

The XSLT stylesheets that we have implemented for the mapping are available from the authors. We encourage other users of CMDI to use and adapt them to their profiles and requirements. Being able to make visible CMDI data in library catalogues contributes to the resources' accessibility. Now, a linguist's entire work (traditional publications *and* research data) can appear in the same catalogue.

Acknowledgments. The NaLiDa project "Zentrum für Nachhaltigkeit linguistischer Daten" 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 `https://openskos.meertens.knaw.nl/ccr/browser`

[U2] The CLARIN Component Registry, see `https://catalog.clarin.eu/ds/ComponentRegistry`

[U3] The Dublin Core Metadata Initiative, see `www.dublincore.org`.

[U4] The 15 DC elements, see `www.dublincore.org/documents/dces`.

[U5] Dublin Core Qualified, see `www.dublincore.org/documents/dcmi-terms`.

[U6] The MARC 21 standard, see `www.loc.gov/marc/bibliographic`.

[U7] The MODS standard, see `www.loc.gov/standards/mods`.

[U8] The METS standard, see `http://www.loc.gov/standards/mets/`.

[U9] The EAD standard, see `https://www.loc.gov/ead/`.

[U10] The Dublin Core to MARC crosswalk, see `http://www.loc.gov/marc/dccross.html`.

[U11] The MARC to Dublin Core crosswalk, see `http://www.loc.gov/marc/marc2dc.html`.

[U12] The Dublin Core to EAD crosswalk, see `http://www.loc.gov/ead/ag/agappb.html#sec3`.

[U13] The MARC to EAD crosswalk, see `http://www.loc.gov/ead/ag/agappb.html#sec4`

[U14] The schema "LexicalResourceProfile", see `https://catalog.clarin.eu/ds/ComponentRegistry?registrySpace=published&itemId=clarin.eu:cr1:p_1290431694579`

6. Bibliographical References

Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A data category registry- and component-based metadata framework. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2010, Malta*. European Language Resources Association.

CLARIN-D. (2012). The clarin-d user guide. `http://media.dwds.de/clarin/userguide/text`.

Godby, C. J., Young, J. A., and Childress, E. (2004). A repository of metadata crosswalks. *D-Lib Magazine*, 10(12). OCLC Online Computer Library Center, Inc, ISSN 1082-9873.

Kemps-Snijders, M., de Bruin, M. J., Kunst, J. P., van der Peet, C. M., Zeeman, R. H. M., and Zhang, J. (2012). Applying "cmdi" in real life: the Meertens case. In *Workshop "Describing Language Resources with Metadata" (LREC-2012)*, Istanbul.

Pierre, M. S. and LaPlant, W. P. (1998). Issues in crosswalking, content metadata standards. Technical report, NISO Standards. `http://www.niso.org/publications/white_papers/crosswalk/`.

Trippel, T. and Zinn, C. (2016). Enhancing the quality of metadata by using authority control. In *5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. Portorož, Slovenia, Colocated with LREC 2016.

Woodley, M. S. (2008). Crosswalks, metadata harvesting, federated searching, metasearching. In M. Baca, editor, *Introduction to Metadata 3.0*. J. Paul Getty Trust, `https://getty.edu/research/publications/electronic_publications/intrometadata/`. ISBN 978-0-89236-966-9 (PDF).

A Metadata of GermaNet in MARC 21 and Dublin Core

Fig. 2 depicts the metadata description of the language resource *GermaNet* in Dublin Core, and Fig. 3 shows this resource expressed in the bibliographic format MARC 21.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<oai_dc:dc xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xmlns:dc="http://purl.org/dc/elements/1.1/"
           xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
           xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                               http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <dc:title>GermaNet</dc:title>
   <dc:title>GermaNet (in OAI): Ein lexikalisch-semantisches Wortnetz</dc:title>
   <dc:coverage>synchron</dc:coverage>
   <dc:date>1997-01-01</dc:date>
   <dc:date>2012-05</dc:date>
   <dc:description>GermaNet ist ein lexikalisch-semantisches Wortnetz, dass Nomina,
       Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische
       Einheiten, die dasselbe Konzept ausdruecken, 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.</dc:description>
   <dc:description>GermaNet is a lexical semantic wordnet describing
       German nouns, verbs, and adjectives. Lexical units expressing the same concept are
       bundled in synsets and the semantic relations between the synsets or the lexical
       units are described.
       GermaNet uses the fundamental principles of the English WordNet and can be seen
       as a light weight ontology or an online thesaurus.</dc:description>
   <dc:description>Diese API ist eine Programmierschnittstelle in JAVA fuer die Nutzung
       der GermaNet-XML-Daten.</dc:description>
   <dc:description>Diese API ist eine Programmierschnittstelle in Perl fuer die Nutzung
       der GermaNet-XML-Daten.</dc:description>
   <dc:description>Der Germanet-Explorer ist eine Software zur Visualisierung der
       semantischen Relationen und lexikalischen Eintraege in GermaNet.</dc:description>
   <dc:description>GernEdiT (GermaNet Editing Tool) ist ein graphischer Editor zum
       Bearbeiten und Erweitern der GermaNet-Datenbank.</dc:description>
   <dc:description>Ein Synset ist in GermaNet die zentrale Repraesentations-
       einheit, in dem lexikalische Einheiten, die dasselbe Konzept ausdruecken,
       zusammengefasst werden.</dc:description>
   <dc:format>text/xml</dc:format>
   <dc:language>Deutsch</dc:language>
   <dc:language>German</dc:language>
   <dc:language>deu</dc:language>
   <dc:publisher>University of Tuebingen</dc:publisher>
   <dc:publisher>Universitaet Tuebingen</dc:publisher>
   <dc:rights>Registrierung erforderlich, eingeschraenkte Nutzung fuer nicht-
       akademische und kommerzielle Nutzung</dc:rights>
   <dc:type>Lexicon</dc:type>
   <dc:type>Wortnetz</dc:type>
</oai_dc:dc>
```

Figure 2: Metadata of GermaNet in Dublin Core.

```
<?xml version="1.0" encoding="utf-8"?>
<record xmlns="http://www.loc.gov/MARC21/slim" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:cmd="http://www.clarin.eu/cmd/" xsi:schemaLocation="http://www.loc.gov/MARC21/slim
        http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
   <leader>      am         3u     </leader>
   <datafield tag="024" ind1="8" ind2=" ">
      <subfield code="a">http://hdl.handle.net/11858/00-1778-0000-0005-896E-B</subfield>
   </datafield>
   <datafield tag="024" ind1="8" ind2=" ">
      <subfield code="a">http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml</subfield>
   </datafield>
   <datafield tag="042" ind1=" " ind2=" ">
      <subfield code="a">cmdi</subfield>
   </datafield>
   <datafield tag="100" ind1="1" ind2="#">
      <subfield code="0">(uri)http://viaf.org/viaf/37069402</subfield>
      <subfield code="0">(uri)http://d-nb.info/gnd/143840657</subfield>
      <subfield code="a">Prof. Dr. Erhard Hinrichs</subfield>
      <subfield code="e">Projektleiter</subfield>
      <subfield code="u">Seminar fuer Sprachwissenschaft, </subfield>
   </datafield>
   <datafield tag="245" ind1="0" ind2="0">
      <subfield code="a">GermaNet</subfield>
      <subfield code="b">Ein lexikalisch-semantisches Wortnetz</subfield>
   </datafield>
   <datafield tag="256" ind1=" " ind2=" ">
      <subfield code="a">74612 Synsets, 99523 Lexikalische Einheiten, 87115 konzeptuelle Relationen,
      3544 lexikalische Relationen, 89819 Number of literals</subfield>
   </datafield>
   <datafield tag="260" ind1="1" ind2="0">
      <subfield code="a">Universitaet Tuebingen</subfield>
      <subfield code="c">1997-01-01</subfield>
   </datafield>
   <datafield tag="270" ind1="1" ind2="0">
      <subfield code="a">Seminar fuer Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tuebingen</subfield>
      <subfield code="d">Deutschland</subfield>
   </datafield>
   <datafield tag="365" ind1=" " ind2=" ">
      <subfield code="b">kostenlos fuer die akademische Nutzung</subfield>
   </datafield>
   <datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">synchron</subfield>
   </datafield>
   <datafield tag="505" ind1="0" ind2=" ">
      <subfield code="a">Java API: Diese API ist eine Programmierschnittstelle in JAVA fuer
                         die Nutzung der GermaNet-XML-Daten.
      </subfield>
   </datafield>
   <datafield tag="505" ind1="0" ind2=" ">
      <subfield code="a">API fuer Perl: Diese API ist eine Programmierschnittstelle in Perl fuer
                         die Nutzung der GermaNet-XML-Daten.
      </subfield>
   </datafield>
   [...]
   <datafield tag="506" ind1=" " ind2=" ">
      <subfield code="a">Registrierung erforderlich, eingeschraenkte Nutzung fuer nicht-
            akademische und kommerzielle Nutzung, kostenlos fuer die akademische Nutzung</subfield>
      <subfield code="u">http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml</subfield>
   </datafield>
   <datafield tag="520" ind1=" " ind2=" ">
      <subfield code="a">GermaNet ist ein lexikalisch-semantisches Wortnetz, dass Nomina,
                  Verben und Adjektive des Deutschen beschreibt. Dabei werden lexikalische
                  [...] oder eine "light-weight ontology" betrachtet werden.</subfield>
   </datafield>
   <datafield tag="653" ind1=" " ind2=" ">
      <subfield code="a">wordnet</subfield>
   </datafield>
   <datafield tag="653" ind1=" " ind2=" ">
      <subfield code="a">lexical resource</subfield>
   </datafield>
   <datafield tag="655" ind1="7" ind2=" ">
      <subfield code="a">Lexicon</subfield>
      <subfield code="2">local</subfield>
   </datafield>
   [...]
   <datafield tag="700" ind1="1" ind2=" ">
      <subfield code="0">(uri)http://d-nb.info/gnd/114724563</subfield>
      <subfield code="0">(uri)http://viaf.org/viaf/17476505</subfield>
      <subfield code="a">Feldweg, Helmut</subfield>
      <subfield code="e">Entwicklung, Annotation</subfield>
   </datafield>
   [...]
   <datafield tag="856" ind1="4" ind2="0">
      <subfield code="u">http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml</subfield>
   </datafield>
</record>
```

Figure 3: Metadata of GermaNet in MARC 21.