EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Seminar für Sprachwissenschaft

Increasing CMDI's Semantic Interoperability with schema.org

Nino Meisinger, Thorsten Trippel, **Claus Zinn**

LREC 2022 Conference, Marseille, 20-25 June 2022

CLARIN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Overview

- Semantic interoperability of CMDI-based metadata based upon single registry
  - metadata fields used in CMDI profiles and components grounded in the CLARIN Concept Registry (CCR)
  - grounding supports CLARIN's Virtual Language observatory to provide semantic search across millions of CMDI-based resources
- But CCR content neither discipline nor CLARIN specific

❖ Why maintain proprietary vocabulary when open, more widely used vocabularies exist elsewhere?

➢ Transformed CMDI-based metadata based on CCR terms into CMDI-based metadata based on *schema.org* terms

➢ Core part of our paper: mapping process

➢ Increase of semantic interoperability !

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Background

- CMDI framework as *de-facto* metadata standard in CLARIN infrastructure
- Framework to provides rich, expressive terms to describe language-related resources and tools FAIR-ly

  - FAIR principles not met by bibliographic standards
  - DublinCore/MARC-21 have no means to describe lexical resources, text or speech corpora, experimental data, tree-banks *etc*.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Component MetaData Infrastructure

- not a metadata schema, but a *framework* to define them
- an ISO standard (ISO 24622-1, ISO 24622-2)
- hierarchical in nature
- *profile* (from which an XML-based schema can be derived from) built from *components* that consist of other components or elementary *elements* (*data categories*)
- elements referenced by IRI, usually resolvable URI
- URI location should contain definition (should point to term registry)
- *values* of data categories can be strings, dates, closed vocabularies (potentially also defined via URIs).

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

CLARIN Concept Registry (CCR)

- started as ISOcat registry, an implementation of the ISO standard ISO 12620:2009)
- Refined to CCR to target only terminological databases (ISO 12620:2019) rather than providing data categories in the more general case
- CCR registry of choice for CMDI metadata designers
- But CMDI specification does not prescribe CCR
- Any term registry or semantic registry can be used to define common ground across CMDI profiles and components

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Bridging Gap to Linked Data

- With CCR being common ground across CMDI profiles, there is little connection to data sources external to CLARIN
- Existing work to convert metadata instances to RDF-based data but this is a syntactic rather than semantic step (Windhouwer et al, 2017)
- Some CMDI components now attach authority file information to person and organisations, *e.g.,* by using GND and VIAF identifiers (Trippel & Zinn, 2020)
- Mapping of data categories to schema.org (Zinn et al., 2012)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

schema.org

- single, light ontology
- backed by major search engines from the beginning
- covers a wide range of topics
- powers Google's Knowledge Graph
- supports various formats, *e.g.*, RDFa, Microdata, JSON-LD
- follows hierarchical type-subtype structure with two building blocks: *types* and *properties*
- every type originates from *Thing* and inherits all properties from its parents
- properties used to describe a type in detail
- well maintained, matured well, has future

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Making use of schema.org: mapping

- Six main CMDI profiles, one for each type of resource
- profiles share all components *non-specific* to the resource
  - *GeneralInfo, Project, Publication, Creation, Documentation, Access, ResourceProxyListInfo*
  - Most elements in these components have equivalent terminology in schema.org
- built tool to convert CMDI-based instances making use of CCR to instances making reference to schema.org.
- built tool to convert XML-based CMDI to JSON-LD based CMDI.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mapping

Field name (in CMDI)	Description Link to DC-based Definition	Definition in schema.org
ResourceName	A short name to identify the language resource CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5	/name
ResourceTitle	The title is the complete title of the resource without any abbreviations CCR_C-2545_d873f2ab-2a2f-29d6-a9ab-260cde57f227	/alternativeHeadline
ResourceClass	Indication of the class, i.e. the type, of a resource CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215	/additionalType
Version	A number that identifies the version of a metadata description a resource or a tool/web service CCR_C-2547_7883d382-b3ce-8ab4-7052-0138525a8ba1	/version
LifeCycleStatus	Indication of the status in the life cycle of a resource. CCR_C-3818_8c4aec73-1654-7565-9575-c4a17425ee29	/creativeWorkStatus
StartYear	The year in which the creation process was started CCR_C-2539_f831f74e-f8ca-4e29-bb02-eb6ca7ea3073	/startDate
CompletionYear	The year in which the creation process was completed CCR_C-2509_3b86afe2-ebde-ba09-8a1c-fe6bdc46a739	/endDate
PublicationDate	The date at which the resource or tool/service was published i.e. announced to the public CCR_C-2538_8b697452-7ef3-9fce-ccf9-a7f344f11317	/datePublished
LastUpdate	The date of the last update CCR_C-2526_979ac535-eaa5-5e59-3cad-51c450234698	/dateModified
TimeCoverage	The time period that the content of a resource is about CCR_C-2502_747eb0cd-03e9-cffb-34cc-d0c8c77e4c5a	/temporalCoverage
LegalOwner	The person or institution who/which holds (all) rights to the resource CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb	/copyrightHolder
Genre	The conventionalized discourse or text types of the content of the resource based on extra-linguistic and internal linguistic criteria CCR_C-2470_d191f2b2-6339-f031-b534-70d526b28357	/genre
FieldOfResearch	Indication of the linguistic field for assigning a resource type to its linguistic context. CCR_C-3796_e89bb008-3e2e-1f70-afa5-e506a6c12683	/about

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Transformation Process (Rules)

1. For each CMD profile, the corresponding schema.org type needs to be specified
   - usually Dataset (default) or SoftwareApplication
2. Each type in the mapping is paired with a JSON-LD context description.
3. then define mappings to the properties of the given type
   - but some CCR entries need to be mapped to a type rather than a property (e.g. licence)
   - and sometimes entire components need to be mapped to a type rather than a property

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

```
<Mappings>
  <Schema.org Type (e.g., DataSet)>
    <Context>JSON-LD Context</Context>
    <Profiles>
      <CMD_Profile_Name>CMD Profile identifier</CMD_Profile_Name>
    </Profiles>
    <Mapping>
      <Property>
        <concept>URL</concept>
        <pattern>XPath</pattern>
        <blacklist>CMD Profile identifier</blacklist>
      </Property>
      <Property>
        [...]
      </Property>
    </Mapping>
  </Schema.org Type (e.g., DataSet)>
  <Schema.org Type (e.g., SoftwareApplication)>
    [...]
  </Schema.org Type (e.g., SoftwareApplication)>
</Mappings>

```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

```
<Mappings>
  <DataSet>
    <Context>"@context": [ "https://schema.org/", {"Component": {"@type": "class","@id": "https://catalog.clarin.eu/ds/ComponentRegistry/#/"}}]</Context>
    <!-- default -->
    <Profiles/>
    <Mapping>
      <id>
        <pattern>/*:CMD/*:Header/*:MdSelfLink</pattern>
      </id>
      <name>
        <concept>http://www.isocat.org/datcat/DC-5428</concept>
        <concept>http://hdl.handle.net/11459/CCR_C-4114_747bf046-1208-940d-36ba-297e4de49e0c</concept>
        <concept>http://purl.org/dc/terms/title</concept> [...]

        <pattern>/*:CMD/*:Components[1]/*/*:GeneralInfo[1]/*:ResourceName[1]</pattern> [...]
        <pattern>/*:CMD/*:Components/*:OLAC-DcmiTerms-ref/*:title</pattern>

        <blacklistProfile>clarin.eu:cr1:p_1527668176124</blacklistProfile> [...]
      </name>
      [...]
      <genre>
        <concept>http://www.isocat.org/datcat/DC-2470</concept>
        <concept>http://hdl.handle.net/11459/CCR_C-2470_d191f2b2-6339-f031-b534-70d526b28357</concept>
        <concept>http://www.isocat.org/datcat/DC-3899</concept>
        <concept>http://hdl.handle.net/11459/CCR_C-3899_c6c608e7-cb2e-1832-09ff-aee36e1f2ed4</concept>
      </genre>
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

```
<license type="CreativeWork">
  <name>
    <concept>URL/concept>
  <name>
</license
```

```
<license type="CreativeWork">
  <name>
    <concept>http://www.isocat.org/datcat/DC-2453</concept>
    <concept>http://www.isocat.org/datcat/DC-2457</concept>
    <concept>http://hdl.handle.net/11459/CCR_C-2457_45bbaa1a-7002-2ecd-ab9d-57a189f694a6</concept>
    <concept>http://www.isocat.org/datcat/DC-3800</concept>
    <concept>http://hdl.handle.net/11459/CCR_C-3800_12a79edd-0ffe-8d82-9831-45d125c54aee</concept>
    <concept>http://www.isocat.org/datcat/DC-5439</concept>
    <concept>http://hdl.handle.net/11459/CCR_C-5439_98bb103d-476a-7f62-54b4-bf9de24d2229</concept>
    <concept>http://hdl.handle.net/11459/CCR_C-5362_8cffd964-f57e-09ed-daed-eeabbf2d22c0</concept>
    <concept>http://purl.org/dc/terms/license</concept>
    <concept>http://purl.org/dc/terms/rights</concept>
  </name>
</license>
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

CMDI metadata based on CCR

```
<cmd:Components>
  <cmdp:ToolProfile>
    <cmdp:GeneralInfo>
      <cmdp:ResourceName xml:lang="en">ProFormA</cmdp:ResourceName>
      <cmdp:ResourceTitle xml:lang="en">ProFormA form based CMDI editing: Core</cmdp:ResourceTitl
      <cmdp:ResourceClass>Tool</cmdp:ResourceClass>
      <cmdp:Version xml:lang="en">0.8</cmdp:Version>
      <cmdp:LifeCycleStatus>released</cmdp:LifeCycleStatus>
      <cmdp:StartYear>2011</cmdp:StartYear>
      <cmdp:CompletionYear>2012</cmdp:CompletionYear>
      <cmdp:PublicationDate>2012-08-10</cmdp:PublicationDate>
      <cmdp:LastUpdate>2012-08-01</cmdp:LastUpdate>
      <cmdp:TimeCoverage/>
      <cmdp:LegalOwner xml:lang="en">SFB 833</cmdp:LegalOwner>
      <cmdp:Genre>other</cmdp:Genre>
      <cmdp:FieldOfResearch>Language Resources</cmdp:FieldOfResearch>
      <cmdp:Location>
        <cmdp:Address>Nauklerstr. 13, 72074 Tübingen</cmdp:Address>
        <cmdp:Region/>
        <cmdp:Country>
          <cmdp:CountryName xml:lang="en">Germany</cmdp:CountryName>
          <cmdp:CountryCoding>DE</cmdp:CountryCoding>
        </cmdp:Country>
      </cmdp:Location>
      <cmdp:Descriptions>
        <cmdp:Description xml:lang="en">ProFormA is a form based editor for CMDI files [...]
        </cmdp:Description>
      </cmdp:Descriptions>
      <cmdp:ModalityInfo> [5 lines]
    </cmdp:GeneralInfo>
```

```
<cmdp:Project>
  <cmdp:ProjectName>SFB 833 INF</cmdp:ProjectName>
  <cmdp:ProjectTitle xml:lang="de">Heterogene Forschungsprimärdat
  <cmdp:ProjectTitle xml:lang="en">Heterogenous Primary Research
  <cmdp:ProjectID>75650358</cmdp:ProjectID>
  <cmdp:Url>http://www.sfb833.uni-tuebingen.de/infrastrukturproje
  <cmdp:Funder>
    <cmdp:fundingAgency>Deutsche Forschungsgemeinschaft (DFG)</cm
  </cmdp:Funder>
  <cmdp:Institution>
    <cmdp:Department xml:lang="de">Sonderforschungsbereich 833: E
  Dynamik und Adaptivität sprachlicher Strukturen</cmdp:Departmer
    <cmdp:Department xml:lang="en">SFB 833: The construction of n
  and adaptivity of linguistic structures</cmdp:Department>
    <cmdp:Url>http://www.sfb833.uni-tuebingen.de/</cmdp:Url>
    <cmdp:Organisation>
      <cmdp:name>Eberhard Karls Universität Tübingen</cmdp:name>
      <cmdp:AuthoritativeIDs>
        <cmdp:AuthoritativeID>
          <cmdp:id>http://viaf.org/viaf/155435537</cmdp:id>
          <cmdp:issuingAuthority>VIAF</cmdp:issuingAuthority>
        </cmdp:AuthoritativeID>
        <cmdp:AuthoritativeID>
          <cmdp:id>http://d-nb.info/gnd/36187-2</cmdp:id>
          <cmdp:issuingAuthority>GND</cmdp:issuingAuthority>
        </cmdp:AuthoritativeID>
        <cmdp:AuthoritativeID>
          <cmdp:id>http://isni.org/isni/0000000121901447</cmdp:id
          <cmdp:issuingAuthority>ISNI</cmdp:issuingAuthority>
        </cmdp:AuthoritativeID>
      </cmdp:AuthoritativeIDs>
    </cmdp:Organisation>
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN


```
{
    "@context": [
        "https://schema.org/",
        {
            "Component": {
                "@type": "class",
                "@id": "https://catalog.clarin.eu/ds/ComponentRegistry/#/"
            }
        }
    ],
    "@type": [
        "SoftwareApplication",
        "clarin.eu:cr1:p_1527668176124"
    ],
    "@id": "https://hdl.handle.net/11022/0000-0007-C5A6-F",
    "name": {
        "@language": "en",
        "@value": "ProFormA"
    },
    "description": {
        "@language": "en",
        "@value": "ProFormA is a form-based editor for CMDI files [...]"
    },
    "url": "https://hdl.handle.net/11022/0000-0007-C5A6-F",
    "identifier": "https://hdl.handle.net/11022/0000-0007-C5A6-F",
    "sameAs": "https://hdl.handle.net/11022/0000-0007-C5A6-F",
    "accessMode": ["other"],
    "dateModified": "2012-08-01",
    "copyrightNotice": [
        {
            "@language": "en",
            "@value": "SFB 833"
        }
    ],
    "genre": ["other"],
    "funder": "Deutsche Forschungsgemeinschaft (DFG)",
    "conditionsOfAccess": [
        "request required",
        {
            "@language": "en",
            "@value": "upon request"
        }
    ],
    "license": {
        "@type": "CreativeWork",
        "name": [
            {
                "@language": "en",
                "@value": "Apache-Licence"
            }
        ]
    },
    "creativeWorkStatus": ["released"],
    "locationCreated": {
        "@type": "Place",
        "address": {
            "@type": "PostalAddress",
            "name": "Nauklerstr. 13, 72074 Tübingen",
            "addressCountry": "DE"
        }
    },
    "creator": [
        {
            "@type": "Organization",
            "@id": "https://viaf.org/viaf/155435537",
            "name": "Eberhard Karls Universität Tübingen"
            "sameAs": [
                "https://viaf.org/viaf/155435537",
                "https://d-nb.info/gnd/36187-2",
                "https://isni.org/isni/0000000121901447"
            ]
        },
        {
            "@type": "Person",
            "@id": "https://d-nb.info/gnd/132884755",
            "givenName": "Thorsten",
            "familyName": "Trippel",
            "sameAs": [
                "https://d-nb.info/gnd/132884755",
                "https://viaf.org/viaf/65179919",
                "https://isni.org/isni/0000000019737791",
                "https://orcid.org/0000-0002-7211-7393"
            ]
        }
    ],
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Discussion: CCR

- The six profiles in our repository share ~ 80% of its CMDI components characterise a resource independently of its specific type.
- the remaining 20% of metadata fields can be used to describe the resource in terms of its specific nature.
- most, if not all, information that is independent of the resource type, can be easily mapped to schema.org vocabulary.
- The situation is different for terminology that describes the nature of a resource type. Here, no satisfying mapping to schema.org vocabulary is possible.
- It is this aspect that shows that the CLARIN concept registry has still an important role to play in the CMDI infrastructure.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Discussion: Conversion

- Conversion into JSON-LD CMDI with no information loss
- Repository now exports legacy CMDI and new CMDI format
- New format is understood outside of the CLARIN world
  - increases findability of resources outside CLARIN community
- Rather than converting CMDI-based instances on the fly, should we write CMDI profiles with *schema.org* vocabulary only and migrate all our instances?
  - But VLO data ingestion tool will need to be informed, otherwise findability in the VLO suffers.
- CCR should focus on terms specific to CLARIN resources