Increasing CMDI’s Semantic Interoperability with schema.org

Nino Meisinger, Thorsten Trippel, Claus Zinn

LREC 2022 Conference, Marseille, 20-25 June 2022
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!
Background

- CMDI framework as *de-facto* metadata standard in CLARIN infrastructure
- Framework to provide 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.*
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).
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
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).
Increasing CMDI's Semantic Interoperability with schema.org – Presentation at LREC 2020, Marseille

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
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.
<table>
<thead>
<tr>
<th>Field name (in CMDI)</th>
<th>Description</th>
<th>Link to DC-based Definition</th>
<th>Definition in schema.org</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResourceName</td>
<td>A short name to identify the language resource</td>
<td>CCR.C-2544.3626545e-a21d-058c-ebfd-241c0464e7e5</td>
<td>/name</td>
</tr>
<tr>
<td>ResourceTitle</td>
<td>The title is the complete title of the resource without any abbreviations</td>
<td>CCR.C-2545.d873f2ab-2a2f-29d6-a9ab-260cde57f227</td>
<td>/alternativeHeadline</td>
</tr>
<tr>
<td>ResourceClass</td>
<td>Indication of the class, i.e. the type, of a resource</td>
<td>CCR.C-3806.e55e9ed6-b099-c21d-a634-3c7f4d22a215</td>
<td>/additionalType</td>
</tr>
<tr>
<td>Version</td>
<td>A number that identifies the version of a metadata description a resource or a tool/web service</td>
<td>CCR.C-2547.7883d382-b3ce-8ab4-7052-0138525a8ba1</td>
<td>/version</td>
</tr>
<tr>
<td>LifeCycleStatus</td>
<td>Indication of the status in the life cycle of a resource.</td>
<td>CCR.C-3818.8c4aec73-1654-7565-9575-c4a17425ee29</td>
<td>/creativeWorkStatus</td>
</tr>
<tr>
<td>StartYear</td>
<td>The year in which the creation process was started</td>
<td>CCR.C-2539.f831f74e-f8ca-4e29-bb02-e6ca7ea3073</td>
<td>/startDate</td>
</tr>
<tr>
<td>CompletionYear</td>
<td>The year in which the creation process was completed</td>
<td>CCR.C-2509.3b86afe2-ebde-ba09-8a1c-fe6bdc46a739</td>
<td>/endDate</td>
</tr>
<tr>
<td>PublicationDate</td>
<td>The date at which the resource or tool/service was published i.e. announced to the public</td>
<td>CCR.C-2538.8b697452-7ef3-9fce-ccf9-a7f344f11317</td>
<td>/datePublished</td>
</tr>
<tr>
<td>LastUpdate</td>
<td>The date of the last update</td>
<td>CCR.C-2526.979ac535-eaa5-5e59-3cad-51c450234698</td>
<td>/dateModified</td>
</tr>
<tr>
<td>TimeCoverage</td>
<td>The time period that the content of a resource is about</td>
<td>CCR.C-2502.747e0cd-03e9-ccfb-34cc-d0c8c77e4c5a</td>
<td>/temporalCoverage</td>
</tr>
<tr>
<td>LegalOwner</td>
<td>The person or institution who/which holds (all) rights to the resource</td>
<td>CCR.C-2956.519a4a9b-2f76-0fd3-090e-f0d6b81a7dbb</td>
<td>/copyrightHolder</td>
</tr>
<tr>
<td>Genre</td>
<td>The conventionalized discourse or text types of the content of the resource based on extra-linguistic and internal linguistic criteria</td>
<td>CCR.C-2470.d191f2b2-6339-f031-b534-70d526b28357</td>
<td>/genre</td>
</tr>
<tr>
<td>FieldOfResearch</td>
<td>Indication of the linguistic field for assigning a resource type to its linguistic context.</td>
<td>CCR.C-3796.e89bb008-3e2e-1f70-afa5-e506a6c12683</td>
<td>/about</td>
</tr>
</tbody>
</table>
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
<Mappings>
   <Schema.org Type (e.g., DataSet)>
      <Context>JSON-LD Context</Context>
      <Profiles>
         <CMD_Profile_Name>CMD Profile identifier</CMD_Profile_Name>
      </Profiles>
   </Schema.org Type (e.g., DataSet)>
   <Schema.org Type (e.g., SoftwareApplication)>
      [...]
   </Schema.org Type (e.g., SoftwareApplication)>
</Mappings>
<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>
      <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-DcmtTerms-ref/*:title</pattern>
    </Mapping>
    <BlacklistProfile>clarin.eu:cr1:p_1527668176124</BlacklistProfile> [...]
  </DataSet>
</Mappings>
Increasing CMDI's Semantic Interoperability with schema.org – Presentation at LREC 2020, Marseille
CMDI metadata based on CCR

ProFormA is a form based editor for CMDI files [...]

Increasing CMDI's Semantic Interoperability with schema.org – Presentation at LREC 2020, Marseille
Increasing CMDI's Semantic Interoperability with schema.org – Presentation at LREC 2020, Marseille
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.
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