Flexibility in Metadata Schemes and Standardisation: the Case of CMDI and DANS Research Data Repositories

Slava Tykhonov; Jerry de Vries; Eko Indarto; Andrea Scharnhorst; Femmy Admiraal; Mike Priddy (DANS-KNAW)

Presentation at ISKO Knowledge Organisation Research Observatory
24 Nov 2021

RESEARCH REPOSITORIES AND DATAVERSE: NEGOTIATING METADATA, VOCABULARIES AND DOMAIN NEEDS
Content

- **Introduction**
  - Who we are
  - CMDI in the ‘wild’ - CLARIN data collections in EASY

- **Creating FAIR metadata and semantic services - case of CMDI Pipeline**
  - Extract-Transform-Load (CMDI) metadata into Dataverse
  - Workflow for linking external concepts to (CMDI) metadata values to make metadata FAIR

- **Lessons learned and pointers**
Introduction

DANS - Royal Netherlands Academy of Arts and Sciences - research data expertise center - long-term preservation - archive (www.easy.dans.knaw; Dataverse.nl; Narcis.nl)

CLARIAH.nl Large Scale Infrastructure Project for Humanities (CLARIN+DARIAH)

DARIAH: Digital Research Infrastructure for the Arts and Humanities

CLARIN: Common Language Resources and Technology Infrastructure

CMDI: Component MetaData Infrastructure: a framework to describe and reuse metadata blueprints

SKOSMOS: Open source web-based SKOS browser and publishing tool

Challenge

How to make the datasets from the CLARIN community in the long-term archive discoverable in the CLARIN infrastructure?

How to search ‘in data’? = How to achieve richer indexing of metadata?
The problem - on a generic level

Communities want to find their **specific** resources – domain specific controlled vocabulary

Platforms and microservices with API’s are means to negotiate between those two perspectives

Archives want to foster cross-domain search and data re-use – rely on generic metadata schemes
We envision a situation where thousands of Dataverse instances (due to EOSC) on the web can be simultaneously search for data.

The old dream of Federated search/Universal catalogue can only be realised if:
(1) Cross-walks; mapping across different metadata schemes are implemented
(2) In metadata schemes we seek for ways to enrich indexes with values from controlled vocabularies

Standard response = standardisation and harmonisation = repository software, certain metadata standards, or certain controlled vocabularies

New response = explore agile solutions (Proof of Concept) which can be implemented by different communities (even smaller ones), so we keep variety and still enable integration.
The problem - on a concrete level
CMDI ‘in the wild’

Typological Database Amsterdam

CLARIN-NL, TDS Curator

Kees Hengeveld

2005

The Typological Database Amsterdam (TDA) focuses on the basic word order and constituent order systems of various languages. Information classifying the speech system of these languages is also provided.

Language and literature studies

This dataset contains CLARIN metadata, i.e., CMDI file(s).

Language typology

TDS Curator

Fixed order of Head & Modifier in referential phrases
Basic Word Order of Head & Modifier in referential phrases
Stem alternation: head in predicator phrases
Copula in property assignment in non-verbal predications
Morphological coding of deviant order HM referential phrase
Morphological coding of deviant order HM Predicative Phrase
Basic Word Order of Head & Modifier in predicate phrases
Stem alternation: head in referential phrases
Content

- **Introduction**
  - Who we are
  - CMDI in the ‘wild’ - CLARIN data collections in EASY

- **Creating FAIR metadata and semantic services - case of CMDI Pipeline**
  - Extract-Transform-Load (CMDI) metadata into Dataverse
  - Workflow for linking external concepts to (CMDI) metadata values to make metadata FAIR

- **Lessons learned and pointers**
Conceptual approach: Semantic interoperability on the infrastructure level - building common solutions for everyone

**Dataverse Semantic API** in release 5.6: [https://github.com/IQSS/dataverse/releases/tag/v5.6](https://github.com/IQSS/dataverse/releases/tag/v5.6)

"Dataset metadata can be retrieved, set, and updated using a new, flatter JSON-LD format - following the format of an OAI-ORE export (RDA-conformant Bags), allowing for easier transfer of metadata to/from other systems (i.e. without needing to know Dataverse's metadata block and field storage architecture). This new API also allows for the update of terms metadata".

**External controlled vocabularies** support is being developed by DANS in SSHOC project and already integrated in Dataverse core in release 5.7.

Proposal: [https://docs.google.com/document/d/1txdcFuxskRx_tLsDQ7KKLFTMR_r9IborDu3V_r445w/](https://docs.google.com/document/d/1txdcFuxskRx_tLsDQ7KKLFTMR_r9IborDu3V_r445w/)

Interfaces: [http://github.com/gdcc/dataverse-external-vocab-support](http://github.com/gdcc/dataverse-external-vocab-support)

Integrations: Wikidata, ORCID, MeSH, Skosmos vocabularies
CMDI Pipeline

- Backbone of our pipeline: Extract-Transform-Load (CMDI) metadata into Dataverse
- One block relevant for semantic services: Mapping across metadata standards
- Another block: Look-up for values in controlled vocabulary registers - enrich indexing
SEMAF: A Proposal for a Flexible Semantic Mapping Framework

March 31, 2021

SEMAF: A Proposal for a Flexible Semantic Mapping Framework

Authors

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>ORCID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broeder, Daan</td>
<td>CLARIN ERIC</td>
<td>0000-0002-8446-3410</td>
</tr>
<tr>
<td>Budroni, Paolo</td>
<td>TU Wien</td>
<td>0000-0001-7490-5716</td>
</tr>
<tr>
<td>Degl’Innocenti, Emiliano</td>
<td>CNR-OVI, ERIHS</td>
<td>0000-0002-3839-9024</td>
</tr>
<tr>
<td>Le Franc, Yann</td>
<td>e-Science Data Factory</td>
<td>0000-0003-4631-148X</td>
</tr>
</tbody>
</table>

This report presents a study for a flexible framework to create, document and publish semantic mappings and cross-walks linking different semantic artefacts within a particular scientific community and across scientific domains. These mappings and cross-walks should be FAIR, as proposed in the FAIR Semantics recommendations. The study draws on the broad expertise of the authors and 25 interviews conducted with community experts. A description for a proposed follow-up implementation project is part of the report.

Proposal: https://zenodo.org/record/4651421#.YT9lyC8RpZI
POC: https://github.com/Dans-labs/semaf-poc
Coming close to the implementation

1. Use Data Catalog Vocabulary (DCAT) mappings for CMDI metadata fields
2. Simple Knowledge Organization System (SKOS) to model a thesauri-like resources with simple skos:broader, skos:narrower and skos:related properties
3. Load CMDI properties and attributes and build a Knowledge Graph out of all elements
4. Enrich the Knowledge Graph with concept URIs from various controlled vocabularies like Skosmos hosted or Wikidata
5. Use different format data-serialization formats suitable for the integration with different systems. For example, json-ld suitable for Dataverse, turtle for Jena Fuseki, RDF for LoD frameworks
Complexity of CMDI is unfolding

After the implementation

- Complexity in CMDI becomes more visible
- Identify core concepts which can be mapped to standard bibliographic schemes as DCAT (red box)
- Possibility to match values of CMDI concepts to other controlled vocabularies (green box)
How does it look when implemented in Dataverse?

Every field can be linked to the appropriate controlled vocabularies in FAIR way!
Greater vision:
Dataverse metadata schemas ingested into a Knowledge Graph

We use SKOS relationships to keep the hierarchy and relationships between metadata fields.
Once in a Knowledge graph: what can we do?

The example of automatic enrichment with Wikidata

Pipeline managed to establish some relationships to Wikidata concepts and automatically updated the dataset with new conceptURIs!
Content

- **Introduction**
  - Who we are
  - CMDI in the ‘wild’ - CLARIN data collections in EASY

- **Creating FAIR metadata and semantic services - case of CMDI Pipeline**
  - Extract-Transform-Load (CMDI) metadata into Dataverse
  - Workflow for linking external concepts to (CMDI) metadata values to make metadata FAIR

- **Lessons learned and pointers**
Lessons learned (I)

Scientific communities and archives have different perspectives on standardisation, and semantic services. In research formalisation (including KOS, ontologies, any ‘model’) is a heuristic device, agile to new research questions, and so intrinsically ‘not interoperable’. In other words, there is a difference between research needs and information needs.
Lessons learned (II)

- We provided a solution for our CMDI problem - by creating a CLARIN compatible Dataverse solution, which via an API can be harvested by the CLARIN search service; we also created another perspective on the CDMI ‘challenge’

- We used Dataverse is a platform due to an open active community;
- The examples we showed you some of are results of a ‘Vision Lab’ - proof of concepts - funded in projects as SSHOC, CLARIAH, EOSC
- The results are envisioned be implemented locally.
- But, in principle the solutions are platform agnostic.
Lessons learned (III)

- In the future, repositories might become nodes in a large searchable knowledge graph and semantic links might enable pathways for contextual/semantic search.

- Part of this future will be automatically supported semantic enrichment at the local instantiations (automatic indexing in a net instead of in an index).

- Problem: keep provenance, authority (trust) - governance between those (micro-service) providers need to be organised. What can we learn from history?
<table>
<thead>
<tr>
<th>References - pointers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CMDI exploration tool</td>
<td><a href="https://github.com/Dans-labs/cmdi-exploration-tool">DANS CMDI converter github</a></td>
</tr>
<tr>
<td>CMDI properties frequency</td>
<td><a href="https://github.com/Dans-labs/VLO-top-profiles">VLO top profiles</a></td>
</tr>
<tr>
<td>CMDI core metadata proposal</td>
<td><a href="https://github.com/Dans-labs/cmdi-core-metadata">Core metadata components design for use cases</a></td>
</tr>
<tr>
<td>DANS CMDI metadata generator</td>
<td>CMDI metadata model published as <a href="https://github.com/Dans-labs/cmdi-core-metadata">TSV files</a> Convertor can extract and show the hierarchy of all fields</td>
</tr>
<tr>
<td>CLARIAH compliant Dataverse Docker module</td>
<td><a href="https://github.com/Dans-labs/cmdi-docker">Dataverse Docker</a> with CMDI metadata schema</td>
</tr>
<tr>
<td>Core metadata components design guidelines</td>
<td><a href="https://github.com/Dans-labs/cmdi-design-guidelines">Guidelines link</a></td>
</tr>
<tr>
<td>Semantic Gateway as plugin app</td>
<td><a href="https://github.com/Dans-labs/smart-datatools">Dataverse gateway</a> <a href="https://github.com/Dans-labs/semantic-gateway-api">Semantic Gateway API</a></td>
</tr>
</tbody>
</table>
References - pointers

Dataverse 5.7 https://github.com/IQSS/dataverse/releases/tag/v5.7

Semantic Gateway: https://github.com/Dans-labs/semantic-gateway

SSHOC task 5.2 http://github.com/SSHOC

SEMAF client https://github.com/Dans-labs/semaf-client

CMDI data model and namespaces: M. Windhouwer, E. Indarto, D. Broeder. CMD2RDF: Building a Bridge from CLARIN to Linked Open Data

Questions?

Slava Tykhonov <vyacheslav.tykhonov@dans.knaw.nl>
Andrea Scharnhorst <andrea.scharnhorst@dans.knaw.nl>