Data Archiving and Networked Services

DANS

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

Clariah

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)

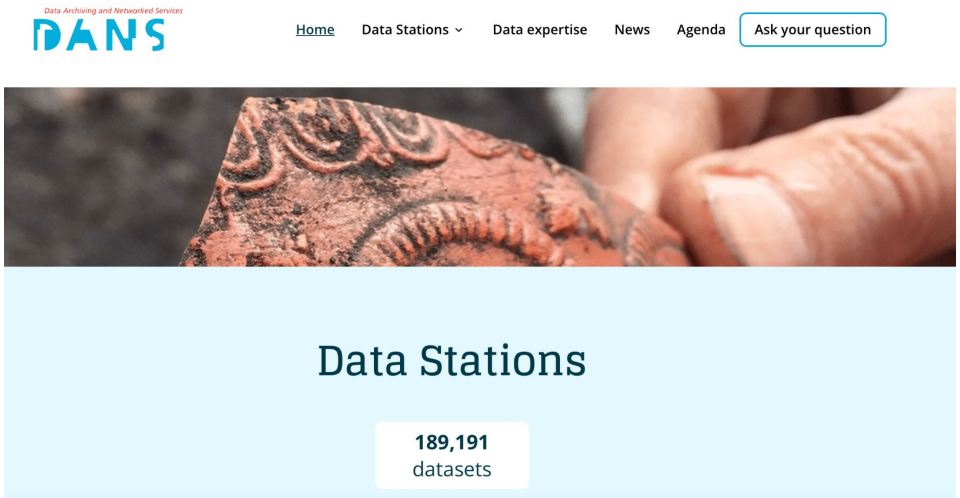
CLARIAH.nl Large Scale Infrastructure Project for Humanities (CLARIN+DARIAH)

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

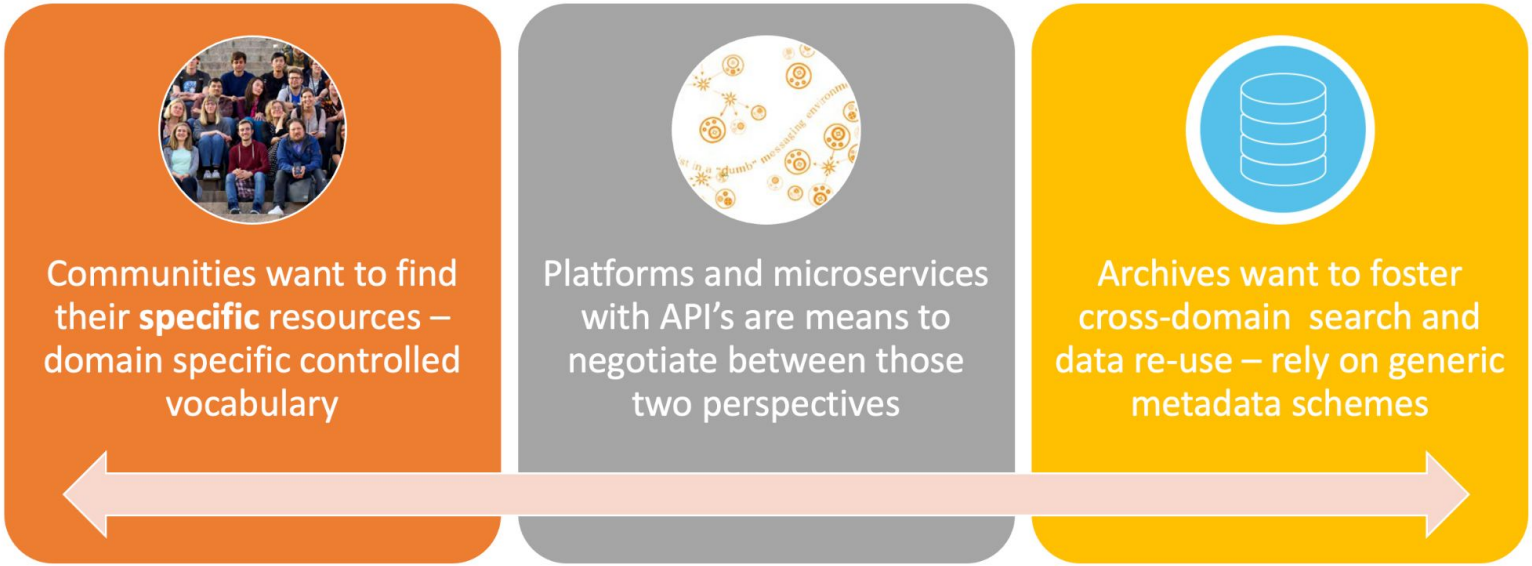


Challenge

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



Vision: Semantic interoperability on the infrastructure level

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’

DANS EASY

HOME REGISTER LOG IN

EASY offers sustainable archiving of research data and access to thousands of datasets.

CLARIN SEARCH > Search help

> Advanced search > Browse

29 RESULTS IN PUBLISHED DATASETS

List Map

Sort by: Choose One

> Typological Database of Intensifiers and Reflexives
Date: 2007 Audience: Language and literature studies
Creators: CLARIN-NL, TDS Curator Access: Open (registered users)
Relevance: 100% relevant Submitted: 2012-04-26

> World Color Survey (summary)
Date: 2005 Audience: Language and literature studies
Creators: CLARIN-NL, TDS Curator Access: Open (registered users)
Relevance: 100% relevant Submitted: 2012-04-26

> Topic Focus Database

REFINE

Published datasets, search: CLARIN ✖

Collections: Common Language Resources and Technology Infra

Search... SEARCH

Advanced search

Audience

Download View details

Dataset Contents / original / Niet-DC-metadata / CMD

- Dataset Contents
  - original
    - Niet-DC-metadata
      - CMDI
    - TDA

	Name
	tda-CMDI-OLAC.xml

TYPOLOGICAL DATABASE AMSTERDAM

‹‹‹ Back to list

Overview | Description | Data files (49)

Persistent identifier	DOI: 10.17026/dans-zhx-a4t2 URN: urn:nbn:nl:ui:13-m1d8-jc
Title	Typological Database Amsterdam
Creator	CLARIN-NL, TDS Curator
Contributor	Kees Hengeveld
Date created (ISO 8601)	2005
Description	The Typological Database Amsterdam (TDA) focuses on the basic word orc constituent order systems of various languages. Information classifying th speech system of these languages is also provided.
Audience	Language and literature studies
Extra CLARIN metadata	This dataset contains CLARIN metadata, i.e., CMDI file(s).
Subject	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 predicate phrases Copula in property assignment in non-verbaal predications Morphological coding of deviant order HM referential phrase Morphological coding of deviant order HM Predicate 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

"*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.

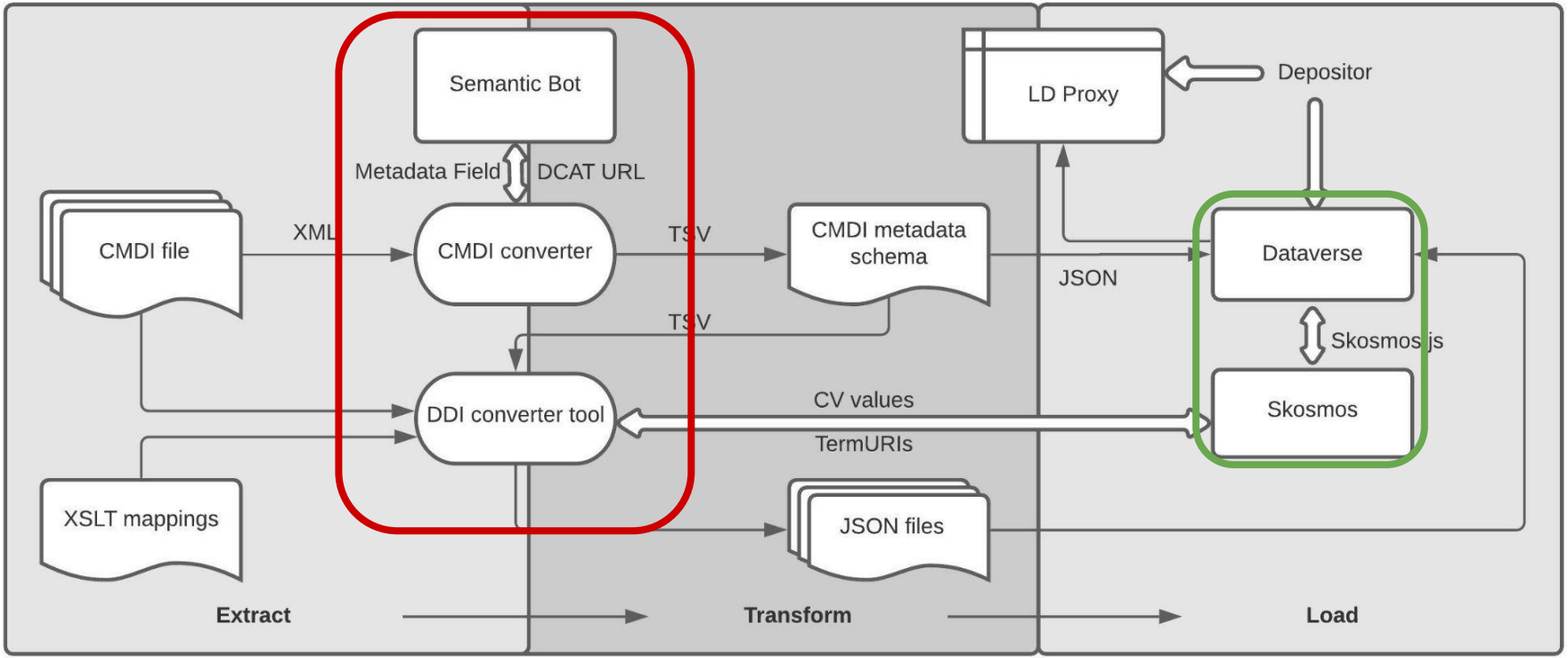
Proposal: https://docs.google.com/document/d/1txdcFuxskRx_tLsDQ7KKLFTMR_r9IBhorDu3V_r445w/

Interfaces: 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

Report | Open Access

SEMAF: A Proposal for a Flexible Semantic Mapping Framework

Broeder, Daan; Budroni, Paolo; Degl'Innocenti, Emiliano; Le Franc, Yann; Hugo, Wim; Jeffery, Keith; Weiland, Claus; Wittenburg, Peter; Zwolf, Carlo Maria

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.

Preview

Page: 1 of 38 Automatic Zoom

SEMAF: A Proposal for a Flexible Semantic Mapping Framework

Version: 1.0, March 2021

Authors

Name	Affiliation	ORCID
Broeder, Daan	CLARIN ERIC	0000-0002-8446-3410
Budroni, Paolo	TU Wien	0000-0001-7490-5716
Degl'Innocenti, Emiliano	CNR-OVI, ERIHS	0000-0002-3839-9024
Le Franc, Yann	e-Science Data Factory	0000-0003-4631-418X

Dans-labs / semaf-poc Public

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

master | 1 branch | 0 tags | Go to file | Add file | Code

4tikhonov Semaf tests for prototyping the conversion of metadata from Dataverse... ee19285 4 days ago 9 commits

semaf	Semaf tests for prototyping the conversion of metadata from Dataver...	4 days ago
semantic-mappings	Semantic mappings folder and sources	18 days ago
sources	Semantic mappings folder and sources	18 days ago
workflow	elasticsearch added to superset infra	18 days ago
LICENSE	Initial commit	18 days ago
README.md	Info how enable Drill connection in Superset	18 days ago
docker-compose.yml	elasticsearch added to superset infra	18 days ago

README.md

semaf-poc

SEMAF Flexible Semantic Mapping Framework Proof of Concept

Presentations and reports

- Flexible Semantic Mapping Framework pdf
- SEMAF final report Zenodo

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


```
<Interviewee>
   <BirthPlace>Veenendaal</BirthPlace>
   <Actor>
      <Role>interviewee</Role>
      <Name>restricted access</Name>
      <FullName>restricted access</FullName>
      <SocialFamilyRole>restricted access</SocialFamilyRole>
      <Age>64</Age>
      <BirthYear>1942</BirthYear>
      <Sex>Male</Sex>
      <Education>Mulo en hulp-etaleur</Education>
      <Profession>restricted access</Profession>
      <Anonymized>true</Anonymized>
      <BirthCountry>
         <Country>
            <Code>NL</Code>
         </Country>
      </BirthCountry>
      <ActorLanguages>
         <ActorLanguage>
            <Language>
               <LanguageName>Dutch</LanguageName>
               <ISO639>
                  <iso-639-3-code>nld</iso-639-3-code>
               </ISO639>
            </Language>
         </ActorLanguage>
      </ActorLanguages>
   </Actor>
</Interviewee>
<Interviewer>
   <Actor>
      <Role>interviewer</Role>
      <Age>53</Age>
      <BirthYear>1954</BirthYear>
      <Sex>Male</Sex>
      <Education>WO</Education>
      <Profession>onderzoeker/publicist</Profession>
      <Anonymized>true</Anonymized>
      <BirthCountry>
         <Country>
            <Code>NL</Code>
         </Country>
      </BirthCountry>
      <ActorLanguages>
         <ActorLanguage>
            <Language>
               <LanguageName>Dutch</LanguageName>
               <ISO639>
                  <iso-639-3-code>nld</iso-639-3-code>
               </ISO639>
            </Language>
         </ActorLanguage>
      </ActorLanguages>
   </Actor>
</Interviewer>
```


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?

Dataverse

Search User Guide Support Dataverse Admin 7

Geographic Unit

Geographic Bounding Box

West Longitude

East Longitude

North Latitude

South Latitude

Unit of Analysis

Mercury, http://skos.um.es/unescothes/C02482

Technical drawing, http://skos.um.es/unescothes/C03980

techni

techni

Technical and vocational education, http://skos.um.es/unescothes/C03978

Technical and vocational study subjects, http://skos.um.es/unescothes/COL150

Technological institutes (Technical colleges), http://skos.um.es/unescothes/C03990

Technical cooperation, http://skos.um.es/unescothes/C03979

Scientific culture (Technical culture), http://skos.um.es/unescothes/C03549

Universe

Time Method

VocabularyURL

Every field can be linked to the appropriate controlled vocabularies in FAIR way!

Greater vision:
Dataverse metadata schemas ingested into a Knowledge Graph

```
@prefix citation: <https://dataverse.org/schema/citation/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

citation: citation:accessToSources [ citation:description "Level of documentation of the original sources." ;
            citation:displayOrder "77" ;
            citation:fieldType "textbox" ;
            citation:metadatablock_id "citation" ;
            citation:name "accessToSources" ;
            citation:title "Documentation and Access to Sources" ] ;
    citation:alternativeTitle [ citation:description "A title by which the work is commonly referred, or an abbreviation of the title." ;
            citation:displayOrder "2" ;
            citation:fieldType "text" ;
            citation:metadatablock_id "citation" ;
            citation:name "alternativeTitle" ;
            citation:title "Alternative Title" ] ;
    citation:alternativeURL [ citation:description "A URL where the dataset can be viewed, such as a personal or project website.  " ;
            citation:displayFormat "<a href=\"#VALUE\" target=\"_blank\">#VALUE</a>" ;
            citation:displayOrder "3" ;
            citation:fieldType "url" ;
            citation:metadatablock_id "citation" ;
            citation:name "alternativeURL" ;
            citation:title "Alternative URL" ;
            citation:watermark "Enter full URL, starting with http://" ] ;
    citation:author [ skos:broader citation:authorAffiliation,
                citation:authorIdentifier,
                citation:authorName ;
            citation:allowmultiples "True" ;
            citation:authorAffiliation [ citation:advancedSearchField "True" ;
                    citation:description "The organization with which the author is affiliated." ;
```

We use SKOS relationships to keep the hierarchy and relationships between metadata fields

```
citation:keyword [ skos:broader citation:keywordValue,
            citation:keywordVocabulary,
            citation:keywordVocabularyURI ;
        citation:allowmultiples "True" ;
        citation:description "Key terms that describe important aspects of the Dataset." ;
        citation:displayOrder "20" ;
        citation:displayoncreate "True" ;
        citation:fieldType "none" ;
        citation:keywordValue [ citation:advancedSearchField "True" ;
```

Compound keyword field with SKOS

Other Dataverse schemas: https://github.com/Dans-labs/semaf-client/tree/cmdi/schema

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!

Files | Metadata | Terms | Versions

Export Metadata

Citation Metadata

Dataset Persistent ID	doi:10.7910/DVN/2Y9VF9
Publication Date	2021-04-04
Title	Code style lookups and class prototypes for ready4 toolkits
Subtitle	Supporting consistent code and automated authoring across the ready4 suite
Alternative URL	https://www.ready4-dev.com/
Other ID	Orygen
Author	Matthew Hamilton (Orygen) - ORCID: https://orcid.org/0000-0001-7407-9194
Contact	Use email button above to contact. Matthew Hamilton (Orygen)
Description	This dataset is a centralised repository of a number of data-files that support the implementation of a consistent code house style and automated code authoring and documentation. It is designed for use when authoring R packages for the ready4 open science framework for modular, replicable and generalisable mental health models.
Subject	Medicine, Health and Life Sciences
Keyword	health-economics, simulation, youth mental health
Language	English
Producer	Orygen https://www.orygen.org.au/
Production Place	Parkville, Australia
Contributor	Project Leader : Matthew Hamilton
Depositor	Hamilton, Matthew
Deposit Date	2021-04-04

Dataverse | Search | User Guide | Support | Sign Up | Log In

Contact	Use email button above to contact. Matthew Hamilton (Orygen)
Description	This dataset is a centralised repository of a number of data-files that support the implementation of a consistent code house style and automated code authoring and documentation. It is designed for use when authoring R packages for the ready4 open science framework for modular, replicable and generalisable mental health models.
Subject	Medicine, Health and Life Sciences
Keyword	Simulation (Wikidata) http://www.wikidata.org/entity/Q20040590 Health Economics, Policy and Law (Wikidata) http://www.wikidata.org/entity/Q15766164 Health Economics (Wikidata) http://www.wikidata.org/entity/Q15679024 Youth mental health services will get overhaul and £1.25bn over five years. (Wikidata) http://www.wikidata.org/entity/Q50593359 Health economics (Wikidata) http://www.wikidata.org/entity/Q69148906 diving (Wikidata) http://www.wikidata.org/entity/Q2252900 Health Economics (Wikidata) http://www.wikidata.org/entity/Q58028918 Simulation of cell rolling and adhesion on surfaces in shear flow: general results and analysis of selectin-mediated neutrophil adhesion. (Wikidata) http://www.wikidata.org/entity/Q34091440 Health economics of dengue: a systematic literature review and expert panel's assessment. (Wikidata) http://www.wikidata.org/entity/Q34592267 Simulation technology for health care professional skills training and assessment. (Wikidata) http://www.wikidata.org/entity/Q33873398 Health economics and cost implications of anxiety and other mental disorders in the United States. (Wikidata) http://www.wikidata.org/entity/Q46140146 Youth mental health services in Italy: An achievable dream? (Wikidata) http://www.wikidata.org/entity/Q47831475 simulation (Wikidata) http://www.wikidata.org/entity/Q45045 Health economics in low income countries: adapting to the reality of the unofficial economy. (Wikidata) http://www.wikidata.org/entity/Q48677723 Health economics (Wikidata) http://www.wikidata.org/entity/Q72012214 Health economics (Wikidata) http://www.wikidata.org/entity/Q5690923 Youth mental health first aid: a description of the program and an initial evaluation (Wikidata) http://www.wikidata.org/entity/Q34586207 Youth mental health interventions via mobile phones: a scoping review. (Wikidata) http://www.wikidata.org/entity/Q38227675 simulation video game (Wikidata) http://www.wikidata.org/entity/Q1610017 Youth mental health in Ireland: a lot done, more to do? (Wikidata) http://www.wikidata.org/entity/Q91302168 Youth mental health in a populous city of the developing world: results from the Mexican Adolescent Mental Health Survey. (Wikidata) http://www.wikidata.org/entity/Q39864096 Youth Mental Health in North Carolina: Creative Innovations in Challenging Times (Wikidata) http://www.wikidata.org/entity/Q90051733 health economics (Wikidata) http://www.wikidata.org/entity/Q31218 Youth mental health reform and early intervention: encouraging early signs. (Wikidata) http://www.wikidata.org/entity/Q51865375

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)

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

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?

References - pointers

CMDI exploration tool	DANS CMDI converter github
CMDI properties frequency	VLO top profiles
CMDI core metadata proposal	Core metadata components design for use cases
DANS CMDI metadata generator	CMDI metadata model published as TSV files Convertor can extract and show the hierarchy of all fields
CLARIAH compliant Dataverse Docker module	Dataverse Docker with CMDI metadata schema
Core metadata components design guidelines	Guidelines link
Semantic Gateway as plugin app	Dataverse gateway Semantic Gateway API
Dataverse metadata schema ingested into Graph	https://github.com/Dans-labs/semaf-client/tree/cmdi/schema

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

Flexible Metadata Schemes for Research Data Repositories. / de Vries, Jerry; Tykhonov, Vyacheslav; Scharnhorst, Andrea; Admiraal, Femmy; Indarto, Eko; Priddy, Mike.
2021. Abstract from CLARIN Annual Conference 2021.
https://www.clarin.eu/content/programme-clarin-annual-conference-2021

Questions?

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