



Flexible metadata schemes for research data repositories
The Common Framework in Dataverse and the CMDI use case

Slava Tykhonov

Jerry de Vries

Eko Indarto

Andrea Scharnhorst

Femmy Admiraal

(DANS-KNAW)

CLARIN annual conference, 27.09.2021

Overall goal for the CMDI core metadata task

The goal mentioned in CMDI strategy 2019-2020: "Ready-made, good quality profiles & components suitable for common use cases and resource types".

DataCite has three types for metadata elements: mandatory, recommended, optional, how to distinguish CMDI core components for different CLARIN centers?

We are part of the specific CMDI task for the design and implementation of CLARIN core metadata components and profiles, and the use of FAIR vocabularies within CLARIN metadata.

Context



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



5 challenges for CMDI

Challenge 1: A proposal of a core set of CMDI metadata as recommendation

Challenge 2: Extraction of CMDI metadata and transform and load the metadata fields into the Dataverse Core set of metadata

Challenge 3: Workflow for prediction and linking concepts from external controlled vocabularies to the CMDI metadata values

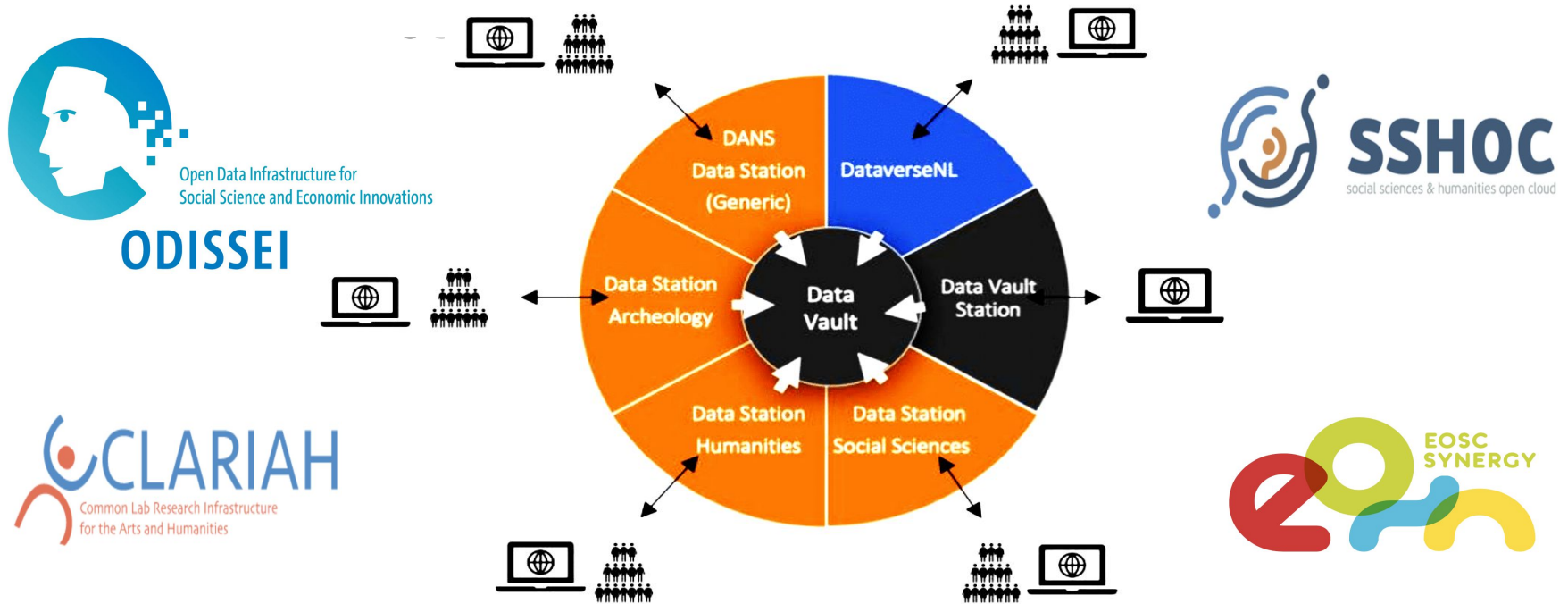
Challenge 4: Extension of the Common Framework with support of FAIR controlled vocabularies to create FAIR metadata

Challenge 5: Extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format

Overall goals for DANS

- CMDI metadata archived in EASY TDR and marked as CLARIN collection with very limited Dublin Core based metadata schema (EDM), curation and further reuse isn't possible
- DANS wants to run Data Station with CMDI metadata schema suitable for different CLARIN use cases
- the long term goal of DANS is to make CMDI datasets harvestable and approachable, and create an interoperability layer with external controlled vocabularies (FAIR Data Point)

DANS Data Stations - Future DANS Data Services



Dataverse is API based data platform and a key framework for Open Innovation!

Semantic interoperability on the infrastructure level

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 will be integrated in Dataverse core in the future releases.

Proposal: https://docs.google.com/document/d/1txdcFuxskRx_tLsDQ7KKLFTMR_r9lBhorDu3V_r445w/

Interfaces: <http://github.com/gdcc/dataverse-external-vocab-support>

Integrations: Wikidata, ORCID, MeSH, Skosmos vocabularies

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, Wirm; 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 Dataverse...	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#.YT9lyC8RpZl>

POC: <https://github.com/Dans-labs/semaf-poc>

Use cases for CMDI metadata

- DDI (DDI)
- Historical documents (HIST)
- Learner corpora (LEARN)
- Lexical/conceptual resources (LEX)
- QUEST ([QUEST](#))
- Software tools/services/components/workflows (SOFT)
- Virtual Collection Registry (VCR)

CMDI example from EASY

```
<?xml version="1.0" encoding="UTF-8"?>
<CMD CMDVersion="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.clarin.eu/cmd/ http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:clp_1369752611610/xsd" xmlns="http://www.clarin.eu/cmd/">
  <Header>
    <MdSelfLink></MdSelfLink>
  </Header>
  <Resources>
    <ResourceProxyList>
    </ResourceProxyList>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <OralHistoryInterviewDANS>
      <ID>IPNV_322</ID>
      <InterviewGeneral>
        <NumberOfSpeakers>2</NumberOfSpeakers>
        <CreationDate>2007-10-05</CreationDate>
        <PublicationDate>2013-03-01</PublicationDate>
        <Duration>01:40:00</Duration>
        <Owner>Veterans Institute, Doorn, The Netherlands</Owner>
        <Genre>interview</Genre>
        <Modality>
          <Modality>Spoken</Modality>
        </Modality>
        <Multilinguality>
          <Multilinguality>Monolingual</Multilinguality>
        </Multilinguality>
        <Access>
          <Availability>All data including audio is accessible for authorised researchers.</Availability>
          <DistributionMedium>Distribution medium will be decided in consultation with Access Contact</DistributionMedium>
          <CatalogueLink>urn:nbn:nl:ui:13-onq-hrv</CatalogueLink>
          <Contact>
            <Address>P.O. Box 93067, 2509 AB Den Haag, The Netherlands</Address>
            <Email>info@dans.knaw.nl</Email>
            <Organisation>Data Archiving and Networked Services (DANS)</Organisation>
            <Telephone>+31 70 3446484</Telephone>
            <Website>www.dans.knaw.nl</Website>
          </Contact>
        </Access>
      </InterviewGeneral>
      <Creators>
        <Creator>
          <Role>Project Manager</Role>
          <Contact>
            <Person>Stef Scagliola</Person>
            <Address>P.O. Box 125, 3940 AC Doorn, The Netherlands</Address>
            <Email>ipnv@veteraneninstituut.nl</Email>
            <Organisation>Veterans Institute</Organisation>
            <Telephone>+31 343 474150</Telephone>
            <Website>www.veteraneninstituut.nl</Website>
          </Contact>
        </Creator>
      </Creators>
      <Location>
        <Address>often stated in the first part of the audio recording</Address>
        <Region>Unknown</Region>
      </Location>
    </OralHistoryInterviewDANS>
  </Components>
</CMD>
```

CMDI exploration tool

develop 2 branches 0 tags

Go to file Add file Code

This branch is 1 commit ahead of master. Pull request Compare

4tkhonov Simple export of fields to TSV format added 2394a83 on Sep 11 27 commits

data	The example of CMDI hierarchy added	7 months ago
tests	CMDI to python dictionary convertor	7 months ago
xml2dict	Simple export of fields to TSV format added	last month
.gitignore	Initial commit	7 months ago
README.md	Example with CMDI folder processing added to Readme	6 months ago
cmdi2dict.py	Simple export of fields to TSV format added	last month

README.md

CMDI/XML exploration tool

by Slava Tykhonov, Data Archiving and Networked Services (DANS-KNAW) <https://dans.knaw.nl>

*This package created for CLARIAH+ WP3 <https://clariah.nl>

Licensed under GPLv3

Usage

```
usage: cmdi2dict.py [-h] [-j] [-s] [-H] [-i inputfile] [-o outputfile] [-D inputfolder]

optional arguments:
  -h, --help            show this help message and exit
  -v, --verbose          verbose mode for debug messages
  -j, --json            convert CMDI format to JSON
```

develop CLARIAH_CMDI / data / CMDIfreq.txt

Go to file ...

4tkhonov Updated with data from CMDI collection Latest commit a17e77b on Mar 31 History

1 contributor

115 lines (115 sloc) 2.03 KB

Raw Blame

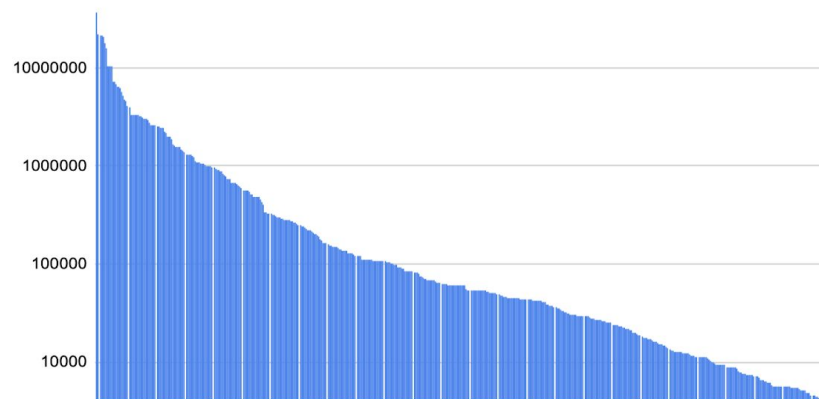
```
1 Keyword 9417344
2 lsp-639-3-code 3891584
3 l50639 1945702
4 LanguageName 1945792
5 TimeInterval 1632640
6 Code 1459200
7 Language 1833696
8 Address 851200
9 InterviewKeywords 755520
10 Organisation 733696
11 Email 733696
12 Telephone 729600
13 Website 729600
14 Country 688000
15 Role 498496
16 Person 498496
17 ActorLanguage 486400
18 Contact 366848
19 MimeType 364800
20 Profession 243200
21 Education 243200
22 Age 243200
23 BirthYear 243200
24 Anonymized 243200
25 ActorLanguages 243200
26 Sex 243200
27 BirthCountry 243200
28 Multilinguality 182400
29 Modality 182400
30 Creator 121728
31 TimeCoverage 121600
32 Compression 121600
33 DistributionMedium 121600
```

Source: [DANS CMDI converter github](https://github.com/dans-cmdl/cmdl-converter)

CMDI properties frequency

	A	B	C	D	E	F
1	XML Property	Frequency (?)		Language		XML property frequency for sample from 20 most common profile names
2	cmdp:Description	36774144		Description		Data gathered from VLO alpha Solr; stats generated using https://github.com/Dans-labs/CLARIAH_CMDI
3	cmdp:AnnotationType	22291456		CMD header		Sample size: 6779 records (1% of 678k files representing 20 most common profile names)
4	cmdp:SizeUnit	21258752				Sample date: 2020-04-08
5	cmdp:Number	21238016				
6	cmdp:iso-639-3-code	20817152				
7	cmdp:Code	18114816				
8	cmdp:MimeType	16102720				
9	cmdp:LanguageName	10574912				
10	cmdp:TotalSize	10572928				
11	cmdp:Name	10415552				
12	cmdp:ISO639	10408576				
13	cmdp:Address	7338496				
14	cmdp:AnnotationFormat	7265280				
15	cmdp:descriptions	6768832				
16	cmd:ResourceRef	6528128				
17	cmd:ResourceType	6528128				
18	edm-componentColor	6286336				
19	cmdp:Country	5667200				
20	cmdp:Language	5295872				
21	cmdp:Continent	4728128				
22	cmdp:Type	4596288				
23	cmdp:Organisation	4050688				
24	cmdp:Email	4046976				
25	cmdp:Compression	3341184				
26	cmdp:NumberOfChannels	3339264				

Histogram (top 500)



Source: [VLO top profiles](#)

Challenge 1: CMDI core metadata proposal

	A	B	C	D	E
1		First 'sprint'			
2	Aspect	VCR	ADP-DDI	Core component properties	Notes
3	Identification	[0-*] Internal identifier [0-1] DOI [0-1] Handle [0-n] Alternate identifier	[0-1] DOI [0-1] Handle [0-*] Internal identifier [0-n] Alternate identifier	[1-n] Identifier <URI> - @type [0-n] Alternate identifier [0-n] Internal identifier	At least one identifier has to be mandatory. Use NOTE: In XML added @cue:atLeastOne="ident" siblings in the same "ident" group should be filled ISSUE: In the editor you can't enter multilingual only the (english) description of the complete co
4	TitleInfo	[1-n] Title [0-n] Subtitle [0-n] Alternative title	[1-n] Title [0-n] Alternative title	[1-n]* Title [0-n] Alternative title [0-n] Subtitle	If "title" is unbounded, I'm not sure I understand makes it different from just another instance of "
5	Description	[1-*] Description	[0-n] Description (abstract)		cardinality 0 doesn't seem right to me, so I went there are a lot of "description" concepts in the co
6	Resource type	[1-1] Resource type [1-1] Resource type general	[1-1] Resource type = data set [0-n] Data kind	[0-n] Identifier [1-n]* Label	
7	Creator	[1-1] Name [0-n] Identifier [0-n] Affiliation [0-1] E-mail address [1-1] Role	[0-n] Creator name	[0-n] Identifier [1-1] Name [0-n]* Alternative name [0-n] Affiliation [0-n] Role [0-1] ContactInfo - [0-n] Email	I have also made the reusable 'ContactInfo' com Type attribute on identifier property?
	VersionInfo	[1-1] Version identifier	[0-n] Version		Merged version into version identifier

Source: [Core metadata components design for use cases](#)

DANS CMDI metadata generator

- CMDI fields generated automatically by our tool from files deposited in EASY
- CMDI metadata model published as [TSV files](#) and uploaded to Dataverse
- Converter can extract and show the [hierarchy of all fields](#):

```
['#document']
[DEBUG] Keys: ['#document']
      /#document
[DEBUG] Keys: ['CMD']
      /#document/CMD
[DEBUG] Keys: ['Header', '#attributes', 'Components', 'Resources']
      /#document/CMD/Header
[DEBUG] Keys: ['MdCreator', 'MdCreationDate', 'MdProfile', 'MdCollectionDisplayName', 'MdSelfLink']
      /#document/CMD/Header/MdCreator
      /#document/CMD/Header/MdCreationDate
      /#document/CMD/Header/MdProfile
      /#document/CMD/Header/MdCollectionDisplayName
      /#document/CMD/Header/MdSelfLink
      /#document/CMD/#attributes
[DEBUG] Keys: ['xmlns:xsi', 'xmlns', 'xsi:schemaLocation', 'CMDVersion']
      /#document/CMD/#attributes/xmlns:xsi
      /#document/CMD/#attributes/xmlns
      /#document/CMD/#attributes/xsi:schemaLocation
      /#document/CMD/#attributes/CMDVersion
      /#document/CMD/Components
[DEBUG] Keys: ['corpusProfile']
      /#document/CMD/Components/corpusProfile
[DEBUG] Keys: ['corpusInfo', 'resourceCommonInfo']
      /#document/CMD/Components/corpusProfile/corpusInfo
[DEBUG] Keys: ['#attributes', 'corpusPartInfo', 'corpusType']
      /#document/CMD/Components/corpusProfile/corpusInfo/#attributes
[DEBUG] Keys: ['ComponentId']
      /#document/CMD/Components/corpusProfile/corpusInfo/#attributes/ComponentId
      /#document/CMD/Components/corpusProfile/corpusInfo/corpusPartInfo
[DEBUG] Keys: ['corpusPartCommonInfo', 'corpusTextInfo', 'mediaType', '#attributes']
      /#document/CMD/Components/corpusProfile/corpusInfo/corpusPartInfo/corpusPartCommonInfo
[DEBUG] Keys: ['geographicCoverageInfo', 'timeCoverageInfo', 'annotationInfo', '#attributes', 'classificationInfo', 'lingualityInfo', 'languageInfo']
      /#document/CMD/Components/corpusProfile/corpusInfo/corpusPartInfo/corpusPartCommonInfo/geographicCoverageInfo
[DEBUG] Keys: ['#attributes', 'geographicCoverage']
      /#document/CMD/Components/corpusProfile/corpusInfo/corpusPartInfo/corpusPartCommonInfo/geographicCoverageInfo/#attributes
[DEBUG] Keys: ['ComponentId']
      /#document/CMD/Components/corpusProfile/corpusInfo/corpusPartInfo/corpusPartCommonInfo/geographicCoverageInfo/#attributes/ComponentId
```

CLARIAH compliant Dataverse Docker module

IQSS / dataverse-docker

Unwatch 17 Star

<> Code Issues 16 Pull requests 2 Actions Projects Wiki Security Insights Settings

clariah 16 branches 3 tags

Go to file Add file Code

This branch is 4 commits ahead, 7 commits behind master. Pull request Compare

4tikhonov CMDI test metadata schema added 01e6380 on Sep 9 192 commits

bridge	add dara plugin	2 years ago
cvmanager	Docker updated	2 years ago
dataversedock	Merge master.	7 months ago
kubernetes	Use newline eof.	2 years ago
metadata	CMDI test metadata schema added	last month
postgresql	fixing commented out line	16 months ago
solr	Master updated	2 years ago
solr7	SOLR schema updated	16 months ago
timbuctoo	Manual extended with extra tips.	2 years ago
README.md	Instruction updated	last month
docker-compose-bridge.yml	Bridge is optional functionality	2 years ago
docker-compose-elst.yml	CVManager support added	2 years ago
docker-compose-local.yml	SSL support add with Traefik 2.0 deployment	last month
docker-compose.yml	SSL support add with Traefik 2.0 deployment	last month
docker-multilingual.yml	Updated to Solr 7.3.0 and Dataverse 4.9.2	2 years ago
initial.bash	Wrong SOLR bug fixed	2 years ago

About

Experiments in running Dataverse in Docker

Readme

Releases 3

Dataverse 5.1.1 with extern... 5 days ago Latest

+ 2 releases

Packages

No packages published Publish your first package

Contributors 9

Languages

- HTML 78.5%
- Shell 7.0%
- Python 6.3%
- Peril 4.2%
- Dockerfile 4.0%

Source: [Dataverse Docker](#) with CMDI metadata schema

Challenge 2: CMDI implementation in Dataverse

Metadata Fields

Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- ☒ Citation Metadata (Required)
- ☐ Geospatial Metadata
- ☐ Social Science and Humanities Metadata
- ☐ Astronomy and Astrophysics Metadata
- ☐ Life Sciences Metadata
- ☐ Journal Metadata
- ☐ CESSDA Metadata Model
- ☒ CMDI Metadata Model

<input checked="" type="checkbox"/> CMDI Geographic Coverage	Required by Dataverse
<input checked="" type="checkbox"/> OralHistoryInterviewDANS	Required by Dataverse
CMDI Geographic Coverage CMDI Country / Nation	Required by Dataverse
OralHistoryInterviewDANS ID	Required by Dataverse
CMDI Geographic Coverage CMDI State/Province	<input type="radio"/> Required <input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> InterviewGeneral	Required by Dataverse
InterviewGeneral NumberOfSpeakers	Required by Dataverse
InterviewGeneral CreationDate	Required by Dataverse

Done

CMDI metadata model in Dataverse

Is it all about
relationships? =>

CMDI Metadata Model ^

OralHistoryInterviewDANS * ?	ID * ?		+
CMDI Geographic Coverage * ?	CMDI Country / Nation * ?	CMDI State/Province ?	+
InterviewGeneral * ?	NumberOfSpeakers * ?	CreationDate * ?	+
	Genre * ?	Owner * ?	
	Multilinguality * ?	PublicationDate * ?	
	Duration * ?	Modality * ?	
Access * ?	Availability * ?		+
InterviewKeyWords * ?			
Country * ?			

Obvious conflict: Dataverse hierarchy limitation, CMDI has no limits of hierarchies

Core metadata components design guidelines

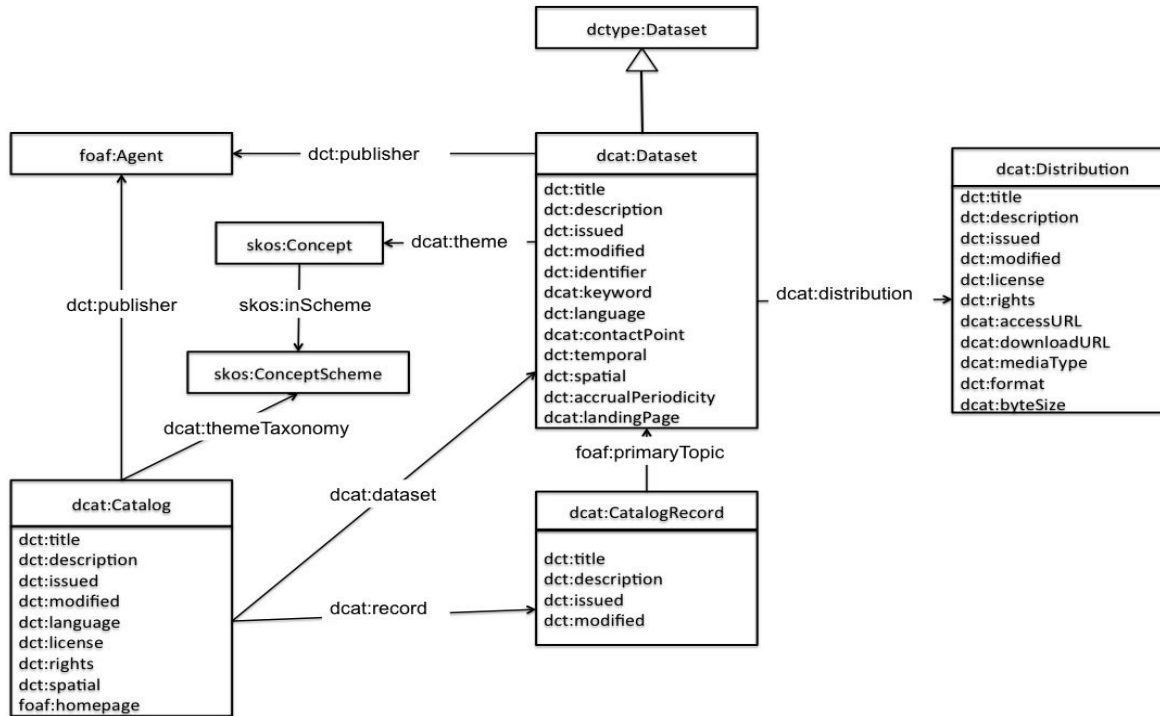
#6	Multilinguality	Element	Allow for multilingual values wherever applicable. Adjust modelling to avoid problems with the unbounded maximal occurrences if necessary.	
#7	Order	Element	Identifier first; then label; then....	Discussion
#9	Information	Component	As a principle, identifier and label (or equivalent property) are present on all 'entity' type components; therefore they should be omitted from 'Info bundle' type	Discussion
#10	Semantics	Component	The semantics of a component must be defined as precisely as necessary by the concept link at the component level . Other approaches to semantic narrowing, such as by means of a 'type' element, should be avoided if the semantic distinction is deemed relevant e.g. for facet mapping in the VLO.	Proposal (TG)
#11	Multilinguality/Cardinality/Documentation	Element	For multilingual elements, intended usage wrt cardinality should be documented, i.e. are multiple values for a single language allowed?	Proposal (TG)
#12	Cardinality/cues	all	For entity type components: include a cue that puts a RECOMMENDED status on identifier and label fields, and at least one field that can help uniquely identify an entity if no identifier can be specified (e.g. an <u>e-mail</u> address or website for a person or organisation)	Proposal (TG)
#14	Display priority	Element	TODO	
#15	Documentation	all	Document all 'soft(er) constraints', even if there are cues for tools	
#16	Documentation	all	Use the 'Documentation templates' table below	
#17	Concept link	all	Use the 'Concept links for common elements' table below	
#18	Concept link	Component	'Entity' type components must have a concept link at the component level.	

Source: [Guidelines link](#)

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

Introduction of Data Catalog Vocabulary (DCAT)



DCAT defines three main classes:

- **dcat:Catalog** represents the catalog
- **dcat:Dataset** represents a dataset in a catalog.
- **dcat:Distribution** represents an accessible form of a dataset

DCAT makes extensive use of terms of RDF, Dublin Core, SKOS, and other vocabs!

Source: [W3C DCAT recommendation](https://www.w3.org/TR/dcat/)

Simple Knowledge Organization System (SKOS)

SKOS models a thesauri-like resources:

- skos:Concepts with preferred labels and alternative labels (synonyms) attached to them (skos:prefLabel, skos:altLabel).
- skos:Concept can be related with skos:broader, skos:narrower and skos:related properties.
- terms and concepts could have more than one broader term and concept.

SKOS allows to create a semantic layer on top of objects, a network with statements and relationships.

A major difference of SKOS is logical “is-a hierarchies”. In thesauri the hierarchical relation can represent anything from “is-a” to “part-of”.



A complex CMDI fragment

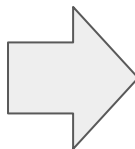
```
<Interviewee>
  <BirthPlace>Veenendaal</BirthPlace>
  <Actor>
    <Role>interviewee</Role>
    <Name>restricted access</Name>
    <FullName>restricted access</FullName>
    <SocialFamilyRole>restricted access</SocialFamilyRole>
    <Age>64</Age>
    <BirthYear>1942</BirthYear>
    <Sex>Male</Sex>
    <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>
```

Some conclusions:

- Top-level concepts (CMDI components) can share the same concepts (called CMDI components)
- Relations between concepts define metadata schema
- Disambiguation of concepts is complicated
- Multilingual components have language indication (for example, keywords in Dutch)
- Hierarchy defined by semantics


Semantics in Dataverse metadata schema


name	facettable	displayoncreate	required	parent	metadatablock	termURI
title	FALSE	TRUE	TRUE		citation	http://purl.org/dc/terms/title
subtitle	FALSE	FALSE	FALSE		citation	
alternativeTitle	FALSE	FALSE	FALSE		citation	http://purl.org/dc/terms/alternative
alternativeURL	FALSE	FALSE	FALSE		citation	https://schema.org/distribution
otherid	FALSE	FALSE	FALSE		citation	
otheridAgency	FALSE	FALSE	FALSE	otherid	citation	
otheridValue	FALSE	FALSE	FALSE	otherid	citation	
author	FALSE	TRUE	FALSE		citation	http://purl.org/dc/terms/creator
authorName	TRUE	TRUE	TRUE	author	citation	
authorAffiliation	TRUE	TRUE	FALSE	author	citation	
authorIdentifierScheme	FALSE	TRUE	FALSE	author	citation	http://purl.org/spar/datacite/AgentIdentifierScheme
authorIdentifier	FALSE	TRUE	FALSE	author	citation	http://purl.org/spar/datacite/AgentIdentifier







*Asterisks indicate required fields


Citation Metadata ^



Title * 


Author * 


Name *  Affiliation 

Identifier Scheme  Identifier 


Contact * 


Name  Affiliation 


E-mail * 

Description * 

This field supports only certain [HTML tags](#).

Text * 

Date 

Subject * 

Dataverse datasetfield API

curl http://localhost:8080/api/admin/datasetfield/title

```
{
  status: "OK",
  - data: {
    name: "title",
    id: 1,
    title: "Title",
    metadataBlock: "citation",
    fieldType: "TEXT",
    allowsMultiples: false,
    hasParent: false,
    controlledVocabularyValues: [ ],
    parentAllowsMultiples: "N/A (no parent)",
    solrFieldSearchable: "title",
    solrFieldFacetable: "title_s",
    isRequired: true
  }
}
```

To do list for Dataverse core:

- add TermURI for metadata fields (DC)
- show external controlled vocabularies available for the specific field
- add multilingual support with 'lang' parameter

We've developed Semantic Gateway as plugin app

Dataverse CVM Setting Generator

Name

Upload your metadata block tsv file:

no file selected

Organisation

<input type="checkbox"/> cessda	<input type="checkbox"/> thesaurus	<input type="checkbox"/> unesco	<input checked="" type="checkbox"/> grid
<input type="checkbox"/> mesh	<input type="checkbox"/> iptc	<input type="checkbox"/> agrovoc	<input type="checkbox"/> faechersystematik

InterviewKeyWords

<input type="checkbox"/> cessda	<input checked="" type="checkbox"/> thesaurus	<input checked="" type="checkbox"/> unesco	<input type="checkbox"/> grid
<input type="checkbox"/> mesh	<input type="checkbox"/> iptc	<input type="checkbox"/> agrovoc	<input type="checkbox"/> faechersystematik


Gateway URL


Dataverse URL

unblock-key


Source: [Dataverse gateway](#)

Semantic Gateway configuration




 main ▾ semantic-gateway / conf / gateway.xml Go to file ...

 **4tikhonov** Configuration updated

Latest commit 19e3440 7 days ago [History](#)

 1 contributor

102 Lines (102 sloc) | 3.34 KB

[Raw](#) [Blame](#)   

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <vocabularies>
3   <ontology name="cessda">
4     <type>cessda</type>
5     <api>https://vocabularies.cessda.eu</api>
6     <uri>v1/suggest/Vocabulary/$vocab/version/1.0/language/en/limit/10/query/$term</uri>
7   </ontology>
8   <ontology name="Unit of Analysis">
9     <type>cessda</type>
10    <vocabulary>Unit of Analysis</vocabulary>
11    <api>https://vocabularies.cessda.eu</api>
12    <uri>v1/suggest/Vocabulary/$vocab/version/1.0/language/en/limit/10/query/$term</uri>
13  </ontology>
14  <ontology name="thesaurus">
15    <type>skosmos</type>
16    <vocabulary>thesaurus</vocabulary>
17    <api>https://vocabulaires.irstea.fr</api>
18    <uri>skosmos/rest/v1/search</uri>
19    <parameters>
20      <vocab>$vocab</vocab>
21      <query>$term</query>
22      <lang>$lang</lang>
23    </parameters>
24  </ontology>
25  <ontology name="wikidata">
26    <type>nde</type>
27    <vocabulary>wikidata</vocabulary>
28    <api>http://demo.netwerkdigitaalrfoed.nl:8080</api>
29    <uri>nde/graphql</uri>
30    <query>query=%20%7B%20terms(match%3A%22$term%22Cdaset%3A%5B%22$vocab%22%5D)%20%7B%20dataset%20terms%20Buri%2C%20altLabel%7D%20%7D%20%7D</query>
31    <parameters>
32      <vocab>$vocab</vocab>
33      <query>$term</query>
```

Dataverse deposit form with connection to ontologies

The screenshot displays the Dataverse deposit form interface. At the top, the Dataverse logo is on the left, and navigation links for Search, User Guide, Support, and Dataverse Admin are on the right. The form is organized into sections, each with a title and a help icon. The 'Geographic Coverage' section includes fields for Country / Nation (with 'Amsterdam' entered), State / Province, City, and Other. The 'Geographic Unit' section has a single text input field. The 'Geographic Bounding Box' section includes fields for West Longitude, East Longitude, North Latitude, and South Latitude. The 'Unit of Analysis' section features a 'Vocabulary' dropdown menu (currently showing 'un|', 'thesaurus', 'grid', and 'agrodoc') and a 'Unit of Analysis Term' field. The 'Universe' section has a large text input field. The 'Time Method' section includes 'Vocabulary' (with 'unesco' entered), 'Time Method Term' (with 'fam' entered), and a 'VocabularyURL' field. Each section has a '+' button to the right of its fields.

Dataverse

Search User Guide Support Dataverse Admin

Geographic Coverage

Country / Nation
Amsterdam

State / Province

City

Other

Geographic Unit

Geographic Bounding Box

West Longitude

East Longitude

North Latitude

South Latitude

Unit of Analysis

Vocabulary
un|
thesaurus
grid
agrodoc

Unit of Analysis Term

Universe

Time Method

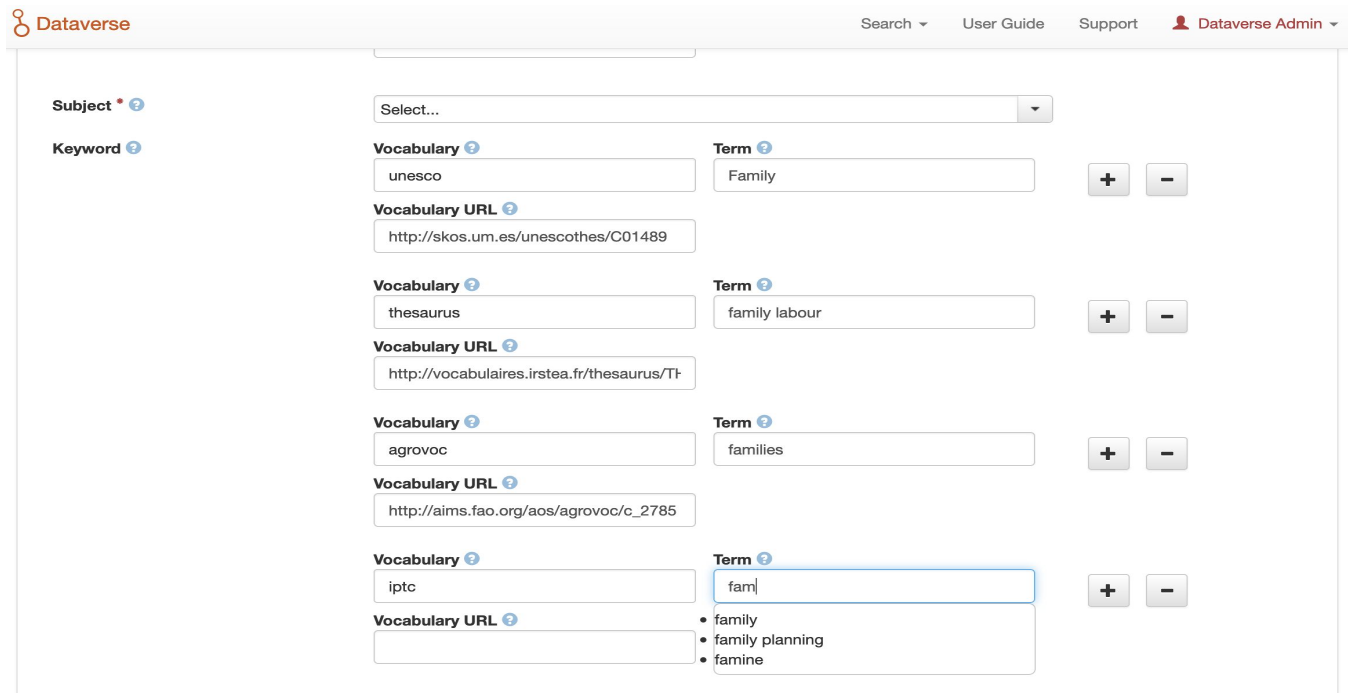
Vocabulary
unesco

Time Method Term
fam

VocabularyURL

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

One metadata field can be linked to many ontologies



The screenshot shows the Dataverse web interface. At the top, there is a navigation bar with the Dataverse logo, a search bar, and links for User Guide, Support, and Dataverse Admin. The main content area is a form for creating or editing a dataset. On the left, there are labels for 'Subject' and 'Keyword'. The 'Subject' field is a dropdown menu. The 'Keyword' field is a text input. Below these, there are four rows of ontology links. Each row consists of a 'Vocabulary' field, a 'Vocabulary URL' field, and a 'Term' field. The 'Vocabulary' field is a dropdown menu. The 'Vocabulary URL' field is a text input. The 'Term' field is a text input with a dropdown arrow. To the right of each 'Term' field are two buttons: a plus sign (+) and a minus sign (-). The first row has 'unesco' in the Vocabulary field, 'http://skos.um.es/unescothes/C01489' in the Vocabulary URL field, and 'Family' in the Term field. The second row has 'thesaurus' in the Vocabulary field, 'http://vocabulaires.irstea.fr/thesaurus/T' in the Vocabulary URL field, and 'family labour' in the Term field. The third row has 'agrovoc' in the Vocabulary field, 'http://aims.fao.org/aos/agrovoc/c_2785' in the Vocabulary URL field, and 'families' in the Term field. The fourth row has 'iptc' in the Vocabulary field, an empty Vocabulary URL field, and 'fam' in the Term field. The 'fam' term is highlighted with a blue border, and a dropdown menu is open below it, showing suggestions: 'family', 'family planning', and 'famine'.

Subject * ?

Keyword ?

Select...

Vocabulary ?

unesco

Vocabulary URL ?

http://skos.um.es/unescothes/C01489

Vocabulary ?

thesaurus

Vocabulary URL ?

http://vocabulaires.irstea.fr/thesaurus/T

Vocabulary ?

agrovoc

Vocabulary URL ?

http://aims.fao.org/aos/agrovoc/c_2785

Vocabulary ?

iptc

Vocabulary URL ?

Term ?

Family

+

-

Term ?

family labour

+

-

Term ?

families

+

-

Term ?

fam

+

-

- family
- family planning
- famine

Language switch in Dataverse will change the language of suggested terms!

The flexibility of Semantic Gateway

Semantic Gateway API 0.1 OAS3

[/openapi.json](#)

Semantic Gateway is Linked Open Data framework for Dataverse.

country

Put this citation in working papers and published papers that use this dataset ✓

namespace

Endpoint to serve namespaces for Controlled Vocabularies ✓

GET `/vocabulary/{term}/` Namespace

default

✓

GET `/configuration/xml` Get Configuration Xml

GET `/configuration/view` Get Configuration Html View

GET `/configuration/edit` Get Configuration Html View

POST `/configuration/edit` Modify Configuration Post

GET `/configuration/download` Download

GET `/` Search

GET `/dv/setting/edit` Get Fields Composer

Source: [Semantic Gateway API](#)

Semantic Gateway lookup API

Scenario: when user selects vocabulary and search for term, API will get filled values and returning back the list of concepts in the standardized format:

GET `/?lang=language&vocab=vocabulary&term=keyword`

examples:

GET [/?lang=en&vocab=unesco&query=fam](#)

GET [/?vocab=mesh&query=sars](#)

Semantic Gateway connected to NDE and SKOSMOS

```
{
  listOfCodes: [
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q103177
      },
      prefLabel: {
        type: "literal",
        value: "severe acute respiratory syndrome"
      }
    },
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q658307
      },
      prefLabel: {
        type: "literal",
        value: "Georg Ossian Sars"
      }
    },
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q1772071
      },
      prefLabel: {
        type: "literal",
        value: "SARS"
      }
    },
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q4408762
      },
      prefLabel: {
        type: "literal",
        value: "Sars River"
      }
    },
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q1772003
      },
      prefLabel: {
        type: "literal",
        value: "structure&activity relationship"
      }
    },
    {
      url: {
        type: "uri",
        value: http://www.wikidata.org/entity/Q84263196
      }
    }
  ]
}
```

Unit of Analysis ?

Vocabulary ?

unesco

Unit of Analysis Term ?

Family disorganization

+

-

VocabularyURL ?

<http://skos.um.es/unescothes/C01483>

Vocabulary ?

thesaurus

Unit of Analysis Term ?

family farms

+

-

VocabularyURL ?

<http://vocabulaires.irstea.fr/thesaurus/Tt>

Vocabulary ?

covidqa

Unit of Analysis Term ?

what are the impacts of COVID-19 amor

+

-

VocabularyURL ?

<https://skosmos.coronawhy.org/COVID1>

Vocabulary ?

covidqa

Unit of Analysis Term ?

mask

+

-

VocabularyURL ?

- what are the best masks for preventing infection by COVID-19?

CMDI data model and namespaces

Default namespace added in Semantic Gateway for CMDI schema to keep all relationships between top-level concepts (metadata fields) in the knowledge graph:

`ns.dataverse.org/cmd_i_component/cmd_i_term`

However, a component or element in CMDI has a unique name among its siblings, so:

```
_ : collection1      a      cmd1 : collection .
_ : actor1           a      cmd2 : Actor .
_ : languages1       a      cmd2 : Actor_Languages .
_ : language1        a      cmd2 : Actor_Languages_Language

_ : collection1      cmdm : contains      _ : actor1 .
_ : actor1           cmdm : contains      _ : languages1 .
_ : languages1       cmdm : contains      _ : language1 .
_ : language1        cmdm : hasElementValue "nld" .
```

Source: [M. Windhouwer, E. Indarto, D. Broeder. CMD2RDF: Building a Bridge from CLARIN to Linked Open Data](#)

Adding component-specific URIs in SKOS

[CMDI Component Registry](#) was created for registered Components/Profiles

Example path in CMDI:

/CMD/Components/corpusProfile/resourceCommonInfo/metadataInfo/metadataCreator/actor
Info/actorType

ns.dataverse.org/cmd1/metadataCreator skos:broader ns.dataverse.org/cmd1/actorInfo

or simply: *cmdi1:metadataCreator skos:related cmdi1:corpusProfile*

CMDI concepts could be linked to the other SKOS concepts on the next step.

How can we link CMDI components in SKOS?

CMDI Component Registry

Loginhelp

Public spaceComponentsFilterNewEdit as newDeleteStatus

Type to filter...Showing 1256 of 1256RSS

Name	Group Name	Domain Name	Creator	Description	Registration Date ▲	Comments
Format	LAT IMDI		Menzo Windhouwer	Format of the published corpus	2016-08-16	0▼
Subject_Languages	LAT IMDI		Menzo Windhouwer	The languages in the corpus that are subject of analysis	2016-08-16	0▼
Document_Languages	LAT IMDI		Menzo Windhouwer	The languages used for documentation of the corpus	2016-08-16	0▼
Content_Languages	LAT IMDI		Alexander	IMDI metadata	2016-08-16	0▼
Subject_Language	LAT IMDI		Menzo Windhouwer	The language in the corpus that is subject of analysis	2016-08-16	0▼
Document_Language	LAT IMDI		Menzo Windhouwer	The language used for documentation of the corpus	2016-08-16	0▼
SL_SignLanguageExperience	LAT IMDI		Alexander	IMDI metadata specific to the sign language profile	2016-08-16	0▼
SL_Deafness	LAT IMDI		Alexander	IMDI metadata specific for the sign language profile	2016-08-16	0▼
SL_Family	LAT IMDI		Alexander	IMDI metadata specific to the sign language profile	2016-08-16	0▼

viewxmlComments (0)

⌵▼

```
<ComponentSpec isProfile="false" CMDVersion="1.2" CMDOriginalVersion="1.2">
  <Header>
    <ID>clarin.eu:crl:c_1407745711995</ID>
    <Name>Document_Languages</Name>
    <Description>The languages used for documentation of the corpus</Description>
    <Status>production</Status>
  </Header>
  <Component name="Document_Languages" CardinalityMin="1" CardinalityMax="1">
    <Component ComponentRef="clarin.eu:crl:c_1407745711994" CardinalityMin="1" CardinalityMax="unbounded"/>
  </Component>
</ComponentSpec>
```

Source: [CMDI Component Registry](#)

Dataverse metadata schema ingested into 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/semaph-client/tree/cmdl/schema>

CMDI conversion to turtle format through Graphs

```
dcterms:creator ns3:Affiliation "Veterans Institute" ;
ns3:Name "Stef Scagliola" ;
cmdi:Address "P.O. Box 125, 3940 AC Doorn, The Netherlands" ;
cmdi:Email "ipnv@veteraneninstituut.nl" ;
cmdi:Telephone "+31 343 474150" ;
cmdi:Website "www.veteraneninstituut.nl" .
```

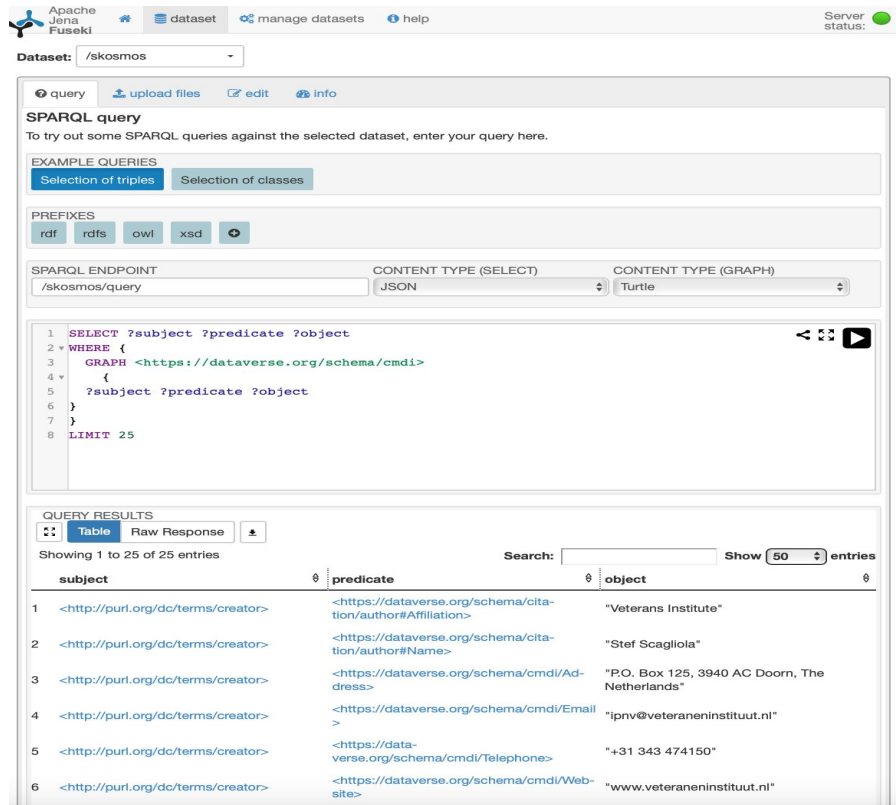
```
<https://dataverse.org/schema/citation/Keyword> ns1:Term [ ns1:Term [ ns1:Term "Contact met thuisfront" ],
  [ ns1:Term "Aanpassingsproblemen" ] ;
  cmdi:TimeInterval "00:00:00-00:10:00" ],
[ ns1:Term [ ns1:Term "Niet praten" ],
  [ ns1:Term "ReÃknie" ],
  [ ns1:Term "Terug in gezinsleven" ],
  [ ns1:Term "Steun van partner" ] ;
  cmdi:TimeInterval "00:10:00-00:20:00" ],
[ ns1:Term [ ns1:Term "PTSS" ],
  [ ns1:Term "Belastende herinneringen" ],
  [ ns1:Term "ReÃknie" ] ;
  cmdi:TimeInterval "00:40:00-00:50:00" ],
[ ns1:Term [ ns1:Term "Nazorg nabestaanden" ],
  [ ns1:Term "Terugkeerreis" ],
  [ ns1:Term "ReÃknie" ],
  [ ns1:Term "Vrijwilligerswerk" ],
  [ ns1:Term "Laatste eer" ],
  [ ns1:Term "Begrafenis militair" ] ;
  cmdi:TimeInterval "00:20:00-00:30:00" ],
[ ns1:Term [ ns1:Term "ReÃknie" ],
  [ ns1:Term "Verwerking" ],
  [ ns1:Term "Nazorg nabestaanden" ],
  [ ns1:Term "Steun van partner" ],
  [ ns1:Term "Veteranenactiviteiten" ] ;
  cmdi:TimeInterval "00:30:00-00:40:00" ],
[ ns1:Term "ReÃknie" ;
  cmdi:TimeInterval "00:50:00-01:00:00" ] .
```

```
cmdi: ns4:document cmdi:CMD ;
cmd:CMD ns4:attributes,
  cmdi:Components,
  cmdi:Header,
  cmdi:Resources ;
#attributes:Components cmdi:OralHistoryInterviewDANS ;
oralhistoryinterviewdans:OralHistoryInterviewDANS cmdi:InterviewAudio,
  cmdi:InterviewContent,
  cmdi:InterviewGeneral,
  cmdi:InterviewMethod,
  cmdi:Interviewee,
  cmdi:Interviewer ;
interviewaudio:InterviewAudio cmdi:SpeechTechnicalMetadata ;
speechtechnicalmetadata:SpeechTechnicalMetadata cmdi:MimeType ;
interviewaudio:InterviewContent cmdi:InterviewSummary,
  cmdi:Mission,
  cmdi:TopicList ;
interviewaudio:InterviewGeneral cmdi:Access,
  cmdi:Creators,
  cmdi:Location,
  cmdi:Modality,
```

Basic workflow:

- using RDFLib to manage namespaces and keep the CMDI hierarchy
- crosswalks to map CMDI fields to DCAT vocabulary
- Serialization to json-ld format allows to deposit CMDI as a dataset in Dataverse
- Converted CMDI records also stored in the triple store

CMDI triples stored in Jena Fuseki triple store



The screenshot shows the Apache Jena Fuseki web interface. At the top, there's a navigation bar with 'dataset', 'manage datasets', and 'help' links. The 'Dataset' dropdown is set to '/skosmos'. Below this, there's a 'query' tab and buttons for 'upload files', 'edit', and 'info'. The main section is titled 'SPARQL query' and includes a text area for entering queries. Below the text area, there are 'EXAMPLE QUERIES' (Selection of triples, Selection of classes) and 'PREFIXES' (rdf, rdfs, owl, xsd). The 'SPARQL ENDPOINT' is '/skosmos/query', 'CONTENT TYPE (SELECT)' is 'JSON', and 'CONTENT TYPE (GRAPH)' is 'Turtle'. The query text area contains the following SPARQL query:

```
1 SELECT ?subject ?predicate ?object
2 WHERE {
3   GRAPH <https://dataverse.org/schema/cmdl>
4   {
5     ?subject ?predicate ?object
6   }
7 }
8 LIMIT 25
```

Below the query text area, there's a 'QUERY RESULTS' section with a 'Table' button and a 'Raw Response' button. The results are displayed in a table with columns 'subject', 'predicate', and 'object'. The table shows 6 entries, with a search bar and a 'Show 50 entries' button.

	subject	predicate	object
1	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/citation/author#Affiliation>	"Veterans Institute"
2	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/citation/author#Name>	"Stef Scagliola"
3	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/cmdl/Address>	"P.O. Box 125, 3940 AC Doorn, The Netherlands"
4	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/cmdl/Email>	"ipnv@veteraneninstituut.nl"
5	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/cmdl/Telephone>	" +31 343 474150"
6	<http://purl.org/dc/terms/creator>	<https://dataverse.org/schema/cmdl/Web-site>	"www.veteraneninstituut.nl"

Benefits:

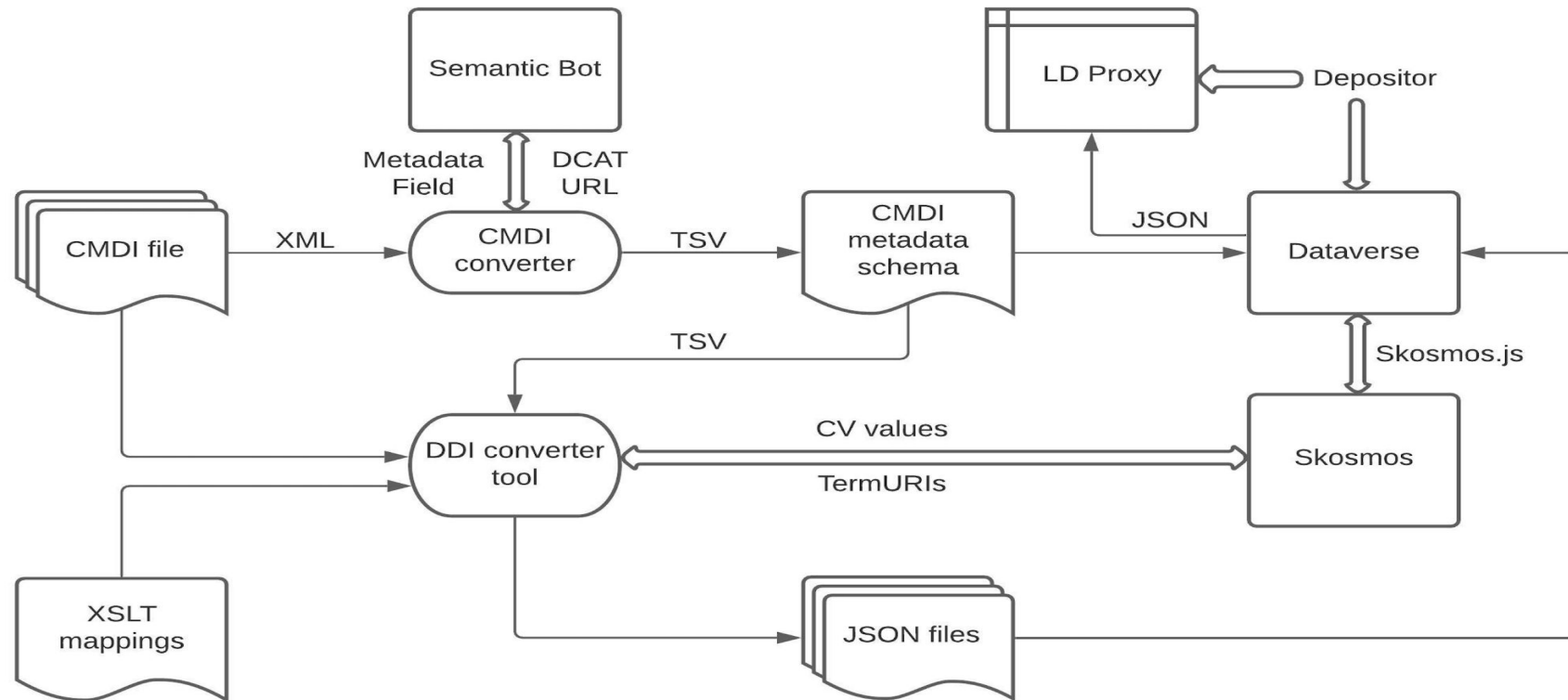
- CMDI Graph stored in RDF with all records fully available as triples
- SPARQL endpoint for complex queries
- Jena has full text search via Lucene and Elasticsearch
- Easy to load and retrieve data from a graph datastore

Challenge 3: Automated workflow to link CMDI concepts to URIs

```
<Language>Dutch</Language>
<InterviewKeyWords>
  <TimeInterval>00:00:00-00:10:00</TimeInterval>
  <Keyword>Bezettingstijd</Keyword>
  <Keyword>Arbeitseinsatz</Keyword>
  <Keyword>Onderduiken</Keyword>
</InterviewKeyWords>
<InterviewKeyWords>
  <TimeInterval>00:10:00-00:20:00</TimeInterval>
  <Keyword>Nasleep WOII</Keyword>
  <Keyword>Militaire dienstplicht</Keyword>
  <Keyword>Ouders en uitzending</Keyword>
  <Keyword>Kennisgeving uitzending</Keyword>
  <Keyword>Militaire keuring</Keyword>
  <Keyword>Kazerneleven</Keyword>
</InterviewKeyWords>
```

- Can we find appropriate FAIR controlled vocabularies for CMDI fields?
- How to resolve disambiguation and link values to URIs?
- Can we link the same value to a few CVs?
- How to link geocodes for geospatial values?

CMDI workflow diagram



Step 1: Querying NDE graphql endpoint

GraphQL IDE interface showing a query and its result.

Query:

```
1 query Terms {
2   terms(sources: ["https://query.wikidata.org/sparql#entities-all"],
3     query: "amsterdam") {
4     source {
5       name
6     }
7     result {
8       __typename
9       ... on Terms {
10        terms {
11          uri
12          prefLabel
13          altLabel
14        }
15      }
16      ... on Error {
17        message
18      }
19    }
20  }
21 }
```

Result:

```
{
  "data": {
    "terms": [
      {
        "source": {
          "name": "Wikidata: alle entiteiten"
        },
        "result": {
          "__typename": "Terms",
          "terms": [
            {
              "uri": "http://www.wikidata.org/entity/Q81888",
              "prefLabel": [
                "AFC Ajax"
              ],
              "altLabel": [
                "Ajax",
                "AFC Ajax Amsterdam",
                "Ajax Amsterdam",
                "Amsterdamsche Football Club Ajax",
                "AFCA"
              ]
            },
            {
              "uri": "http://www.wikidata.org/entity/Q22062165",
              "prefLabel": [
                "Amsterdam"
              ],
              "altLabel": []
            },
            {
              "uri": "http://www.wikidata.org/entity/Q3614694",
              "prefLabel": [
                "Amsterdam"
              ],
              "altLabel": []
            },
            {
              "uri": "http://www.wikidata.org/entity/Q26938585",
              "prefLabel": [
                "Amsterdam"
              ],
              "altLabel": []
            },
            {
              "uri": "http://www.wikidata.org/entity/Q4748822",
              "prefLabel": [
                "Amsterdam"
              ],
              "altLabel": []
            }
          ]
        }
      }
    ]
  }
}
```

QUERY VARIABLES **REQUEST HEADERS**

Step 2: Querying Skosmos API with python module

Skosmos API

The Skosmos REST API is a read-only interface to the data stored on the vocabulary server. The URL namespace is the base URL of the Skosmos instance followed by `/rest/v1/`.

Most methods return the data as UTF-8 encoded JSON-LD, served using the `application/json` MIME type. The data consists of a single JSON object which includes JSON-LD context information (in the `@context` field) and one or more fields which contain the actual data. Some methods (`data`) return other formats (RDF/XML, Turtle, RDF/JSON) with the appropriate MIME type.

The API supports Cross-Origin Resource Sharing by setting the Access-Control-Allow-Origin HTTP header to `"*"` for all requests.

The API supports the JSONP convention of appending a callback parameter to any URL. The returned data will then be wrapped in a JavaScript function call using the function name provided as the callback parameter value. JSONP wrapped data will be served using the `application/javascript` MIME type.

Global methods

Show/Hide | List Operations | Expand Operations

Vocabulary-specific methods

Show/Hide | List Operations | Expand Operations

Concept-specific methods

Show/Hide | List Operations | Expand Operations

GET

`/vocabulary/data`

RDF data of the whole vocabulary or a specific concept. If the vocabulary has support for it, MARCXML data is available for the whole vocabulary in each language.

Parameters

Parameter	Value	Description	Parameter Type	Data Type
vocid	<input type="text" value="(required)"/>	a Skosmos vocabulary identifier e.g. "stw" or "yso"	path	string
format	<input type="text"/>	The MIME type of the serialization format, e.g "text/turtle" or "application/rdf+xml". If not specified, HTTP content negotiation (based on the Accept header) is used to determine a suitable	query	string

pip install skosmos-client

Step 3: collecting concepts in the common Pandas dataframe

	uri	prefLabel	altLabel	hiddenLabel	scopeNote	broader	narrower	related	keyword
0	http://www.wikidata.org/entity/Q2388514	[inlichtingen analyse]							Inlichtingen
1	http://www.wikidata.org/entity/Q16069934	[Inlichtingen voor duivenliefhebbers]							Inlichtingen
2	http://www.wikidata.org/entity/Q105202100	[Inlichtingenbureau]			[Nederlandse stichting]				Inlichtingen
3	http://www.wikidata.org/entity/Q47913	[inlichtingendienst]							Inlichtingen
4	http://www.wikidata.org/entity/Q60469229	[Inlichtingendienst Buitenland]							Inlichtingen
0	http://www.wikidata.org/entity/Q194489	[arbeidsloon]	[loon, bezoldigd, bezoldiging, maandgeld, sala...						Salaris
1	http://www.wikidata.org/entity/Q2640178	[loonschaal]	[barema, salarischaal, salarisgroep, salariskl...						Salaris
2	http://www.wikidata.org/entity/Q1344886	[loonstrook]	[loonstrookje, salarisafrekening, salarisstroo...						Salaris
3	http://www.wikidata.org/entity/Q33963	[Salar]			[taal]				Salaris
4	http://www.wikidata.org/entity/Q64374595	[Salaris]			[achternaam]				Salaris
5	http://www.wikidata.org/entity/Q1879195	[salaris van één dollar]							Salaris
6	http://www.wikidata.org/entity/Q19764471	[Salariskloof fors gegroeid]			[Wikinieuws-artikel]				Salaris
7	http://www.wikidata.org/entity/Q1106776	[salarisplafond]			[algemeen]				Salaris
8	http://www.wikidata.org/entity/Q1369344	[salarisplafond]			[in de sportwereld]				Salaris
9	http://www.wikidata.org/entity/Q51126965	[Salarisverhoging ING-topman Hamers gaat toch ...			[Wikinieuws-artikel]				Salaris

Step 4: Filtering out geospatial concepts



Main page
Community portal
Project chat
Recent changes
Random item
Query Service
Nearby
Help
Donate

Geographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Edit this page

Item Discussion

Read View history

Search Wikidata

Yugoslavia (Q36704)

1918–1992 country in Southeastern and Central Europe
yu | Jugoslavija

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Yugoslavia	1918–1992 country in Southeastern and Central Europe	yu Jugoslavija
Spanish	Yugoslavia	Antiguo país del Sureste de Europa que existió entre 1918 y 1992	
Catalan	lugoslàvia	antic estat d'Europa (1918-1992)	
Galician	lugoslavia	No description defined	lugoslavia - Југославија

All entered languages

Statements

instance of	historical country	
	start time	4 February 2003
	▼ 0 references	
sovereign state	end time	27 April 1992
	▼ 0 references	
Mediterranean country	start time	1 December 1918 <i>Gregorian</i>

Wikipedia (130 entries)

af	Joego-Slawië
als	Jugoslawien
am	ዩጋስላቪያ
ang	Geugoslafia
ar	يوغوسلافيا
ary	يوجوسلافيا
arz	يوجوسلافيا
ast	Yugoslavia
az	Yugoslaviya
bar	Jugoslawien
ba	Югославия
be_x_old	Югаславія
be	Югаславія
bg	Югославия
bn	যুগোস্লাভিয়া
br	Yugoslavia
bs	Jugoslavija
ca	lugoslàvia
cbk_zam	Yugoslavia
ckb	یوگوسلافیا
cs	Jugoslávie
cu	Югославия
cy	Iwgoslafia
da	Jugoslavien
de	Jugoslawien
dsb	Jugosławjarska
el	Γιουγκοσλαβία

Step 5: Caching concepts and recognizing geocodes

CMDI fragment:

```
<InterviewKeyWords>
  <TimeInterval>02:50:00-03:00:00</TimeInterval>
  <Keyword>Individuele uitzending</Keyword>
  <Keyword>Trots</Keyword>
  <Keyword>Medaille-uitreiking</Keyword>
</InterviewKeyWords>
<InterviewSummary>
  <TimeInterval>00:00:00-03:00:00</TimeInterval>
  <Summary>De rode draad in het verhaal van deze vrouwelijke veteraan is haar
individuele uitzending naar Kosovo eind 1999. Ze beschrijft de moeilijkheden die een
individueel uitgezonden militair kan tegengekomen. Ze geeft impressies van de
levensomstandigheden van de bevolking van Kosovo vlak na oorlog. De geïnterviewde vertelt ook
over haar opleiding en dienstdtijd bij de Koninklijke Luchtmacht en wat het voor haar betekent
om 'veteraan' te zijn. </Summary>
</InterviewSummary>
<TopicList>
</TopicList>
<Mission>
  <SpatialCoverage>Kosovo</SpatialCoverage>
  <TimeCoverage>1999-2000</TimeCoverage>
</Mission>
```



```
"Kosovo": {
  "nde": {
    "SpatialCoverage": "Kosovo",
    "geo": {
      "geocode": "XXK",
      "uri": "http://www.wikidata.org/entity/Q1246"
    }
  },
  "rawdata": {
    "data": {
      "terms": [
        {
          "result": {
            "__typename": "Terms",
            "terms": [
              {
                "altLabel": [
                  "Autonome Provincie Kosovo en Metohija (1990-1999)"
                ],
                "broader": [],
                "hiddenLabel": [],
                "narrower": [],
                "prefLabel": [
                  "Autonome Provincie Kosovo en Metohija"
                ],
                "related": [],
                "scopeNote": [
                  "1990-1999"
                ],
                "uri": "http://www.wikidata.org/entity/Q1255"
              },
              {
                "altLabel": [
                  "UCK",
                  "Kosovaarse Bevrijdingsleger",
                  "Kosovo bevrijdingsleger",
                  "U\u00c7K"
                ],
                "broader": [],
                "hiddenLabel": [],
                "narrower": [],
                "prefLabel": [
                  "Bevrijdingsleger van Kosovo"
                ],
                "related": [],
                "scopeNote": [],
                "uri": "http://www.wikidata.org/entity/Q193366"
              }
            ]
          }
        }
      ]
    }
  }
}
```


Step 6. Keeping all CMDI locations in dataframe

	cmdi	cmdi location	country	ISO_3166-1_alpha-3	termURI
0	./cmdi/easy-dataset:42421_easy-file:4297253_IP...	Voormalig Joegoslavië	Jugoslavija		http://www.wikidata.org/entity/Q36704
1	./cmdi/easy-dataset:42192_easy-file:4298646_IP...	Libanon	Republic of Lebanon	LBN	http://www.wikidata.org/entity/Q822
2	./cmdi/easy-dataset:35869_easy-file:4298051_IP...	Nederlands-Indië	Netherlands East Indies		http://www.wikidata.org/entity/Q188161
3	./cmdi/easy-dataset:41954_easy-file:4298559_IP...	Nederlands-Indië	Netherlands East Indies		http://www.wikidata.org/entity/Q188161
4	./cmdi/easy-dataset:46684_easy-file:4297372_IP...	Kosovo	Republic of Kosovo	XKX	http://www.wikidata.org/entity/Q1246
...
724	./cmdi/easy-dataset:46758_easy-file:4297615_IP...	Afghanistan	Islamic Republic of Afghanistan	AFG	http://www.wikidata.org/entity/Q889
725	./cmdi/easy-dataset:36002_easy-file:4298729_IP...	Europa	Europe		http://www.wikidata.org/entity/Q458
726	./cmdi/easy-dataset:35990_easy-file:4298530_IP...	Nederlands-Indië	Netherlands East Indies		http://www.wikidata.org/entity/Q188161
727	./cmdi/easy-dataset:35968_easy-file:4298371_IP...	Europa	Europe		http://www.wikidata.org/entity/Q458
728	./cmdi/easy-dataset:42303_easy-file:4298839_IP...	Voormalig Joegoslavië	Jugoslavija		http://www.wikidata.org/entity/Q36704

Enriched metadata published in Dataverse

IPNV_842

Draft Unpublished



2021, "IPNV_842", <https://doi.org/10.5072/FK2/KYF9SZ>, Root, DRAFT VERSION ?

Cite Dataset ▾

[Learn about Data Citation Standards.](#)

Publish Dataset

Edit Dataset ▾

Contact Owner

Share

Dataset Metrics ?

0 Downloads ?

Files

Metadata

Terms

Versions

Add + Edit Metadata

Citation Metadata ^

Dataset Persistent ID ?

doi:10.5072/FK2/KYF9SZ

Title ?

IPNV_842

Name ?

Stef Scagliola

Affiliation ?

(Veterans Institute)

Text ?

De geïnterviewde heeft als oorlogsvrijwilliger bij de Koninklijke Luchtmacht gediend in Nieuw Guinea. Eerst volgde hij een opleiding tot automonteur en diende hij een aantal jaren bij het 306 Squadron in Duitsland. De geïnterviewde vertelt over de legering aan boord onderweg naar Nieuw Guinea. De legering was slecht en de verveling aan boord was heel erg. Hij vertelt over de werkomstandigheden na aankomst op Biak en spreekt over de geldzaken tijdens de uitzending. De geïnterviewde had wel contact met de Papoea's. Hij vindt het jammer dat hij niet meer voor hen had kunnen doen. Hij vertelt over de goede contacten met een vriend en de commandant tijdens de uitzending.

Term ?

Klimatologische omstandigheden

Deposit Date ?

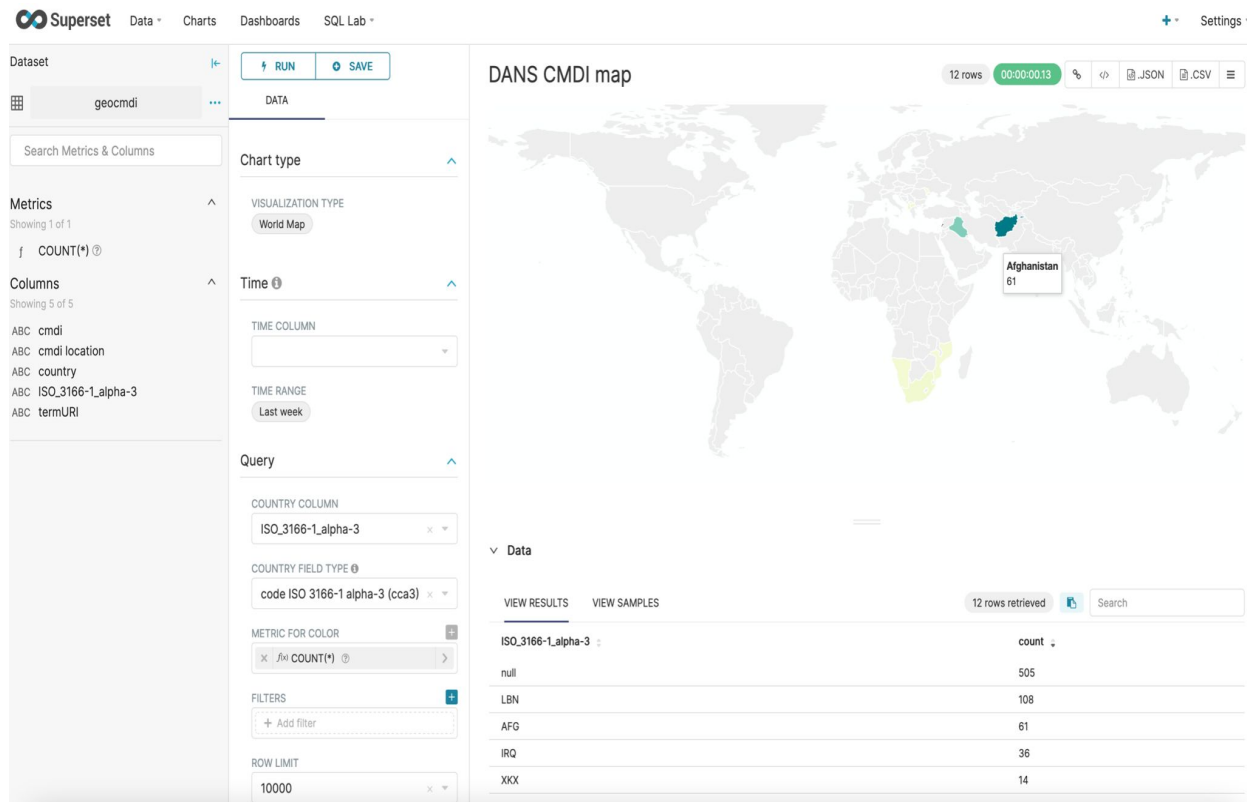
2009-10-16

- We're not using XSLT or any other transformations!
- All operations on metadata enrichment are directly in the Knowledge Graph
- Dataverse Semantic API allows to import json-ld representation of dataset

Data analysis and CMDI visualizations with Superset

- Apache Superset integration with Dataverse was created in SSHOC project
- Superset allows to connect to various external data sources or databases (postgresql, mysql, SOLR, MongoDB, ...). Every new connection considered as a separate dataset
- about 50 different visualizations available out-of-the-box in Superset (charts, maps, wordclouds, network graphs, ...)
- Dataverse with ingested CMDI metadata could be connected to Superset directly

Visualizing existent countries from CMDI on the map with Superset



We can visualize structured CMDI geospatial metadata enriched with iso codes

Only existent countries supported!

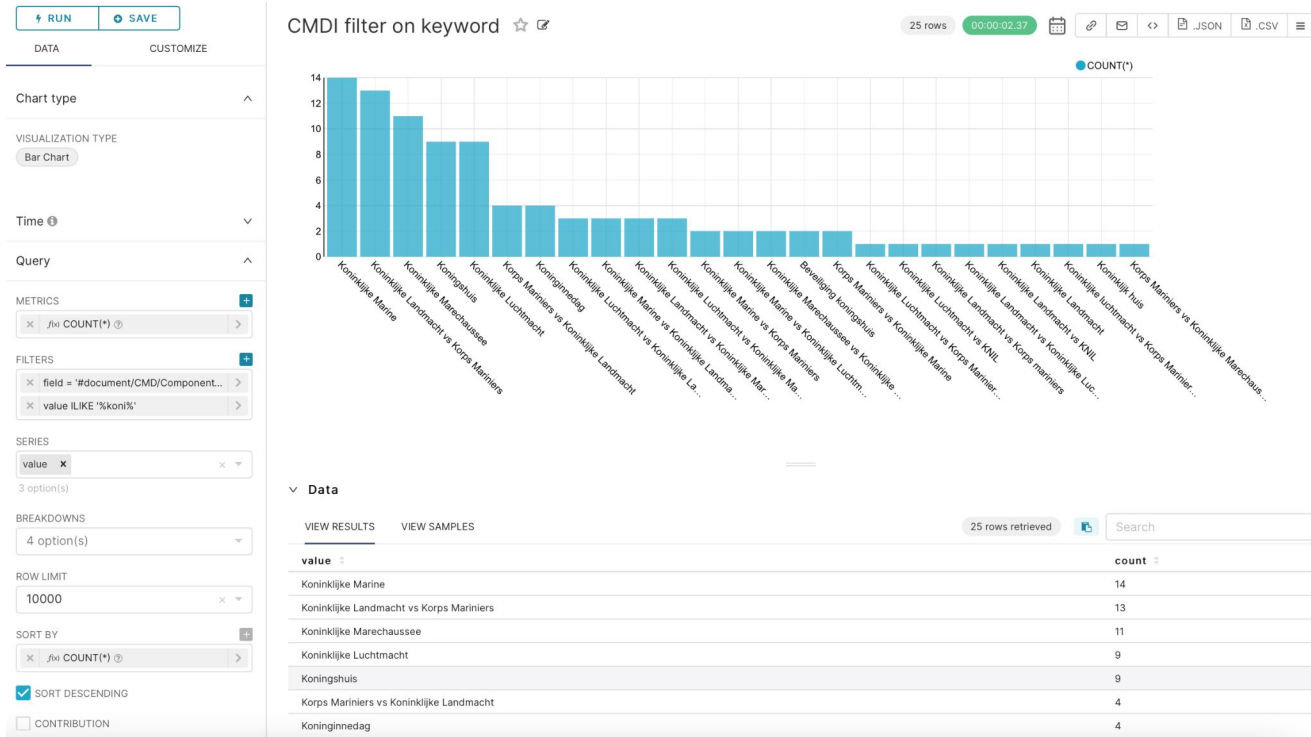
Map can be published as a widget or created on dashboard

Geospatial overview of DANS CMDI datasets

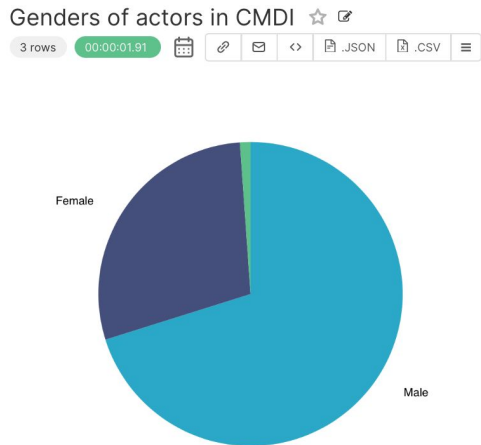
country	COUNT(*)
Netherlands East Indies	204
Jugoslavija	133
Europe	112
Republic of Lebanon	108
Islamic Republic of Afghanistan	61
Kingdom of Cambodia	38
Jumhūriyyat al-'Irāq	36
Republic of Kosovo	14
Island of Cyprus	9
République d'Haïti	4
Republika e Shqipërisë	3
Former Yugoslav Republic of Macedonia	2
Republic of South Africa	2
Republic of Namibia	1
Republic of Moldova	1
República de Moçambique	1



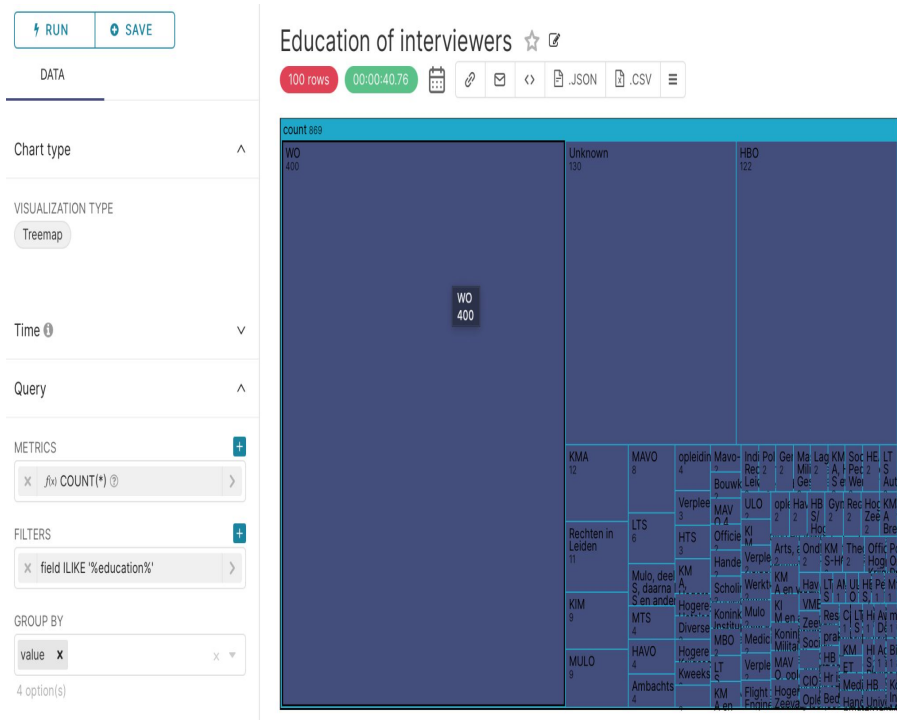
CMDI metadata explorations



Getting new insights is easy!



Analyzing literal values in CMDI Graph



Challenge 4: CMDI extension with support of CVs for FAIR metadata

- We can predict appropriate vocabularies for every field after the running automated pipeline
- CMDI metadata schema could be extended with new CVs support in Dataverse
- one CMDI field can be linked to a few CVs in the same time (for example, Geospatial field to Wikidata and Geonames). User can select appropriate CV in the deposit form during the metadata creation or edit process.
- We're adding a configurable process to add or remove CVs in Dataverse

Suggestions for the usage of FAIR CVs

- Dutch Digital Heritage Network <https://netwerkdigitaalerfgoed.nl>
- Skosmos instances, for example, <https://bartoc-skosmos.unibas.ch/en/>

Skosmos client to access vocabularies <https://pypi.org/project/skosmos-client/>

- ORCID API to link CMDI records to identifiers of researchers
<https://info.orcid.org>
- CESSDA CV Service <https://vocabularies.cessda.eu>

More are coming!

<https://github.com/CLARIAH/awesome-humanities-ontologies>

Example of the CV configuration in Dataverse

```
[{
  "field-name": "cvocDemo",
  "term-uri-field": "cvocDemoTermURI",
  "cvoc-url": "https://skosmos.dev.finto.fi/",
  "js-url": "/resources/js/skosmos.js",
  "protocol": "skosmos",
  "retrieval-uri": "https://skosmos.dev.finto.fi/rest/v1/data?uri={0}",
  "term-parent-uri": "",
  "languages": "en, fr, es, ru",
  "vocabs": {
    "unesco": "http://skos.um.es/unescothes/CS0000"
  },
  "managed-fields": {
    "vocabularyName": "cvocDemoVocabulary",
    "termName": "cvocDemoTerm",
    "vocabularyUri": "cvocDemoVocabularyURI"
  },
  "retrieval-filtering": {
    "@context": {
      "termName": "https://schema.org/name",
      "vocabularyName": "https://dataverse.org/schema/vocabularyName",
      "vocabularyUri": "https://dataverse.org/schema/vocabularyUri",
      "lang": "@language",
      "value": "@value"
    },
    "@id": {
      "pattern": "{0}",
      "params": ["@id"]
    },
    "termName": {
      "pattern": "{0}",
      "params": ["/graph/uri=@id/prefLabel"]
    },
    "vocabularyName": {
      "pattern": "{0}",
      "params": ["/graph/type=skos:ConceptScheme/prefLabel"]
    },
    "vocabularyUri": {
      "pattern": "{0}",
      "params": ["/graph/type=skos:ConceptScheme/uri"]
    }
  }
},
-]
```

Configuration in pluggable JavaScript:

- Field cvocDemo connected to “unesco” controlled vocabulary hosted by Skosmos
- 4 languages available (en, fr, es, ru)
- js-url pointing to javascript gateway to read and transform output from external API endpoint
- every Skosmos concept cached internally in Dataverse to increase the sustainability

Challenge 5: export from Dataverse metadata back to CMDI

Basic requirements:

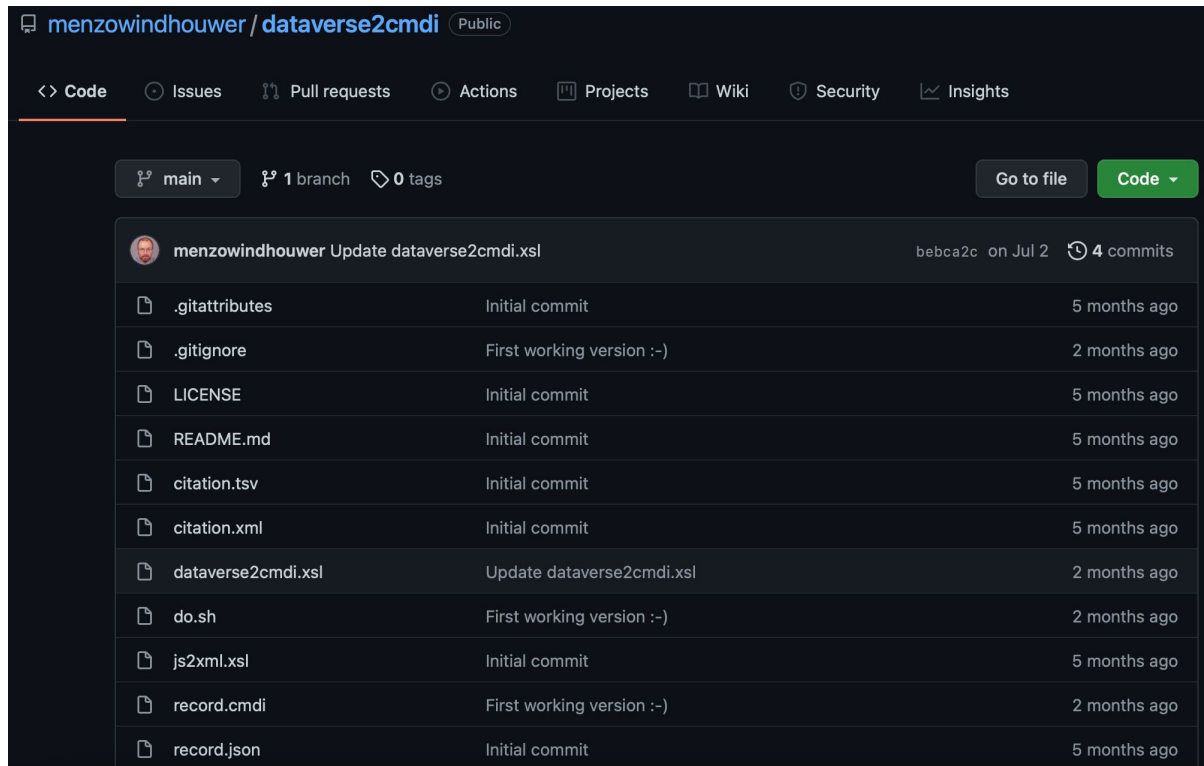
Dataverse metadata schema should have CMDI metadata that can be extended by custom components used by CLARIN centers in the different countries.

Original relationships between fields and concepts should be kept, custom components should be added to SKOS schema.

Users should be able to download metadata in the original CMDI format without losing quality.

This work is in progress...

Dataverse2cmdi export tool



menzowindhouwer / dataverse2cmdi Public

<> Code Issues Pull requests Actions Projects Wiki Security Insights

main 1 branch 0 tags Go to file Code

menzowindhouwer Update dataverse2cmdi.xsl bebca2c on Jul 2 4 commits

.gitattributes	Initial commit	5 months ago
.gitignore	First working version :-)	2 months ago
LICENSE	Initial commit	5 months ago
README.md	Initial commit	5 months ago
citation.tsv	Initial commit	5 months ago
citation.xml	Initial commit	5 months ago
dataverse2cmdi.xsl	Update dataverse2cmdi.xsl	2 months ago
do.sh	First working version :-)	2 months ago
js2xml.xsl	Initial commit	5 months ago
record.cmdi	First working version :-)	2 months ago
record.json	Initial commit	5 months ago

Tips:

- developed by KNAW HuC in the collaboration with DANS
- Intended for the conversion from Dataverse json-ld back to cmdi
- using Semantic Mappings to restore fields from Dataverse citation block

Questions?

Slava Tykhonov (DANS-KNAW)

Jerry de Vries (DANS-KNAW)

Andrea Scharnhorst (DANS-KNAW)

Eko Indarto (DANS-KNAW)

Femmy Admiraal (DANS-KNAW)

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

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