

*Data Archiving and Networked Services*



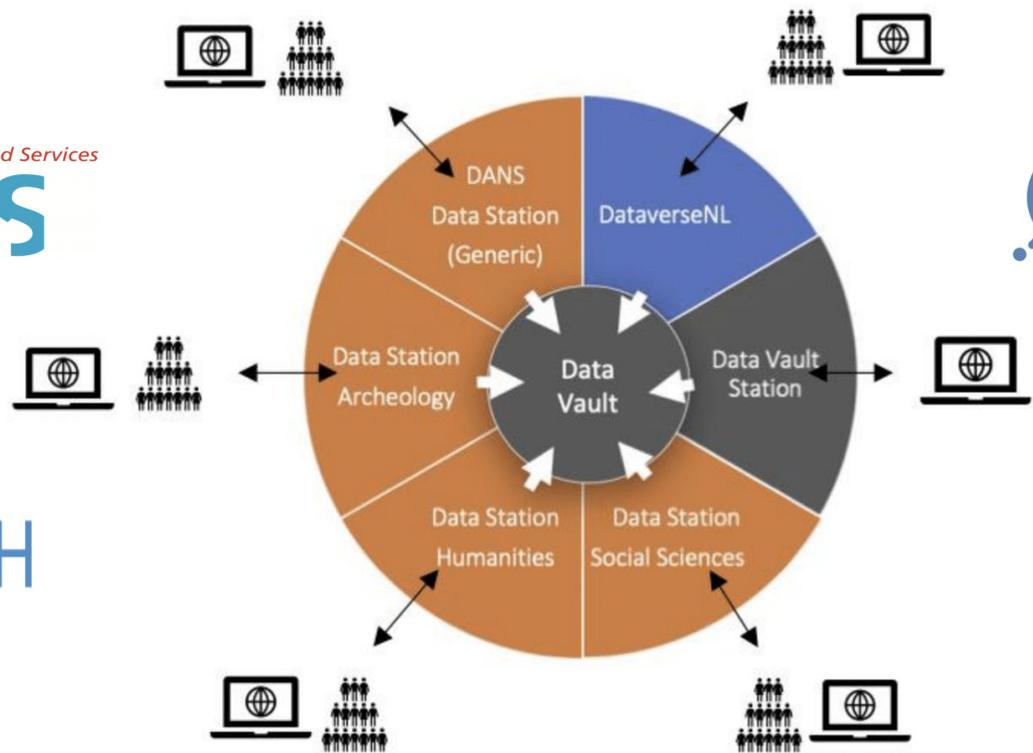
# Knowledge Graphs and Semantic Search in Dataverse

Dataverse Community call, Harvard IQSS

Slava Tykhonov, R&D  
(DANS-KNAW, the Netherlands)

22 August 2023

# DANS Data Stations - Future Data Services



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

# 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 and will form shared Data Lake.

The *old dream* of Federated search/Universal catalogue can only be realised if:

- (1) Crosswalks; 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 (**centralized**) = standardisation and harmonisation = repository software, certain metadata standards, or certain controlled vocabularies

New response (**distributed**) = explore agile solutions (Proof of Concepts) which can be implemented by different communities (even smaller ones), so we keep variety and still enable integration in the Distributed Data Network by applying Linked Data technologies, and building global Knowledge Graph.

# “Archive in a box” features (SSHOC Dataverse)

- Dockerized version of Dataverse application and shared networked services
- **fully automatic Dataverse** deployment with Traefik proxy
- Dataverse configuration managed through environmental file .env
- different **Dataverse distributions** with services on your preference suitable for different use cases and research communities
- **external controlled vocabularies** support (demo of CESSDA CMM metadata fields connected to Skosmos framework)
- S3 compatible MinIO storage support for Cloud Storage
- data previewers integrated in the Dataverse distribution
- startup process managed through scripts located in init.d folder
- automatic SOLR reindex
- external services integration with PostgreSQL triggers
- support of custom metadata schemes (CESSDA CMM, CLARIN CMDI, ...)
- built-in Web interface localization uses [Dataverse language pack](#) to support **multiple languages** out of the box

<https://github.com/IQSS/dataverse-docker>

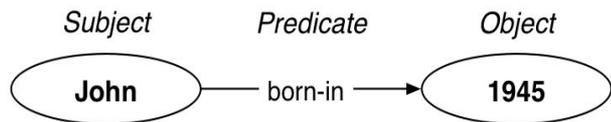
# What is knowledge graph and how it can help?

There is no single commonly accepted definition of a knowledge graph. Most definitions view the topic through a Semantic Web lens and include these features:<sup>[11]</sup>

- *Flexible relations among knowledge in topical domains*: A knowledge graph (i) defines **abstract classes** and relations of entities in a schema, (ii) mainly describes real world entities and their interrelations, organized in a graph, (iii) allows for potentially interrelating arbitrary entities with each other, and (iv) covers various topical domains.<sup>[12]</sup>
- *General structure*: A network of entities, their semantic types, properties, and relationships.<sup>[13][14]</sup>
- *Supporting reasoning over inferred ontologies*: A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.<sup>[2]</sup>

Source: Wikipedia

# Semantic triple and Resource Description Framework (RDF)

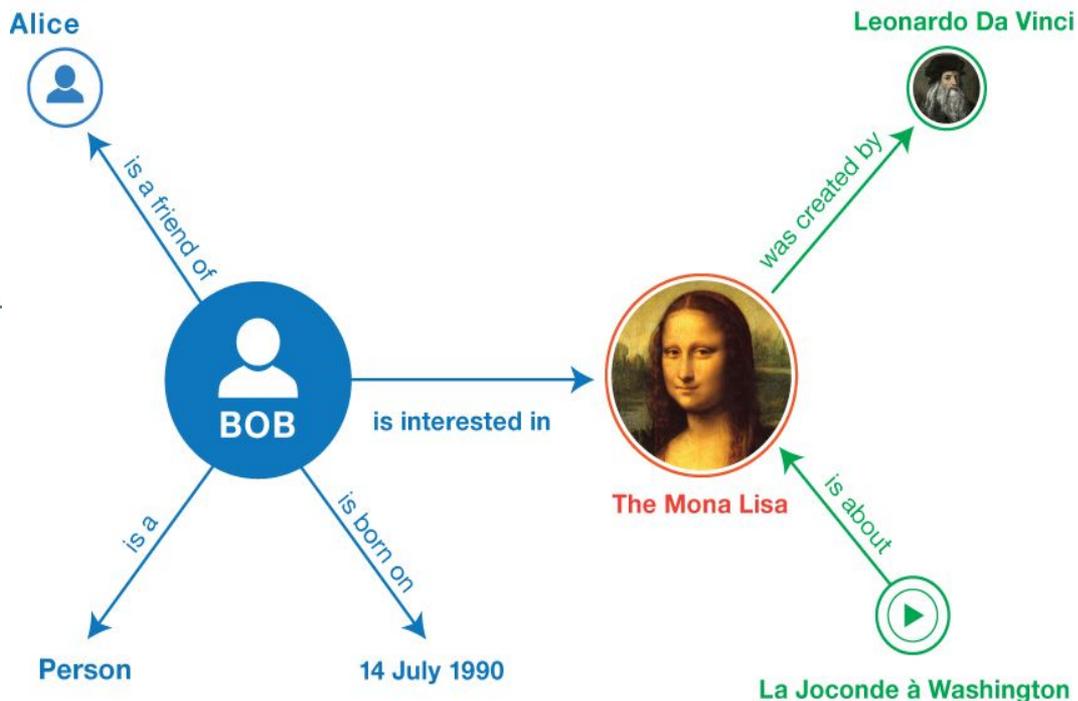


“This format enables knowledge to be represented in a machine-readable way. Particularly, every part of an RDF triple is individually addressable via unique URIs — for example, the statement “Bob knows John” might be represented in RDF as:

```
http://example.name#BobSmith12
http://xmlns.com/foaf/0.1/knows
http://example.name#JohnDoe34.
```

Given this precise representation, semantic data can be unambiguously queried and reasoned about.”

Wikipedia: [Semantic triple](#)



# Dataverse Metadata Schema in Semantics with TermURI

#metadataBlock	name							
	citation							
#datasetField	name	allowmultiples	facetable	displayoncreat	required	parent	metadatablock_	termURI
	title	FALSE	FALSE	TRUE	TRUE		citation	<a href="http://purl.org/dc/terms/title">http://purl.org/dc/terms/title</a>
	subtitle	FALSE	FALSE	FALSE	FALSE		citation	
	alternativeTitle	FALSE	FALSE	FALSE	FALSE		citation	<a href="http://purl.org/dc/terms/alternative">http://purl.org/dc/terms/alternative</a>
	alternativeURL	FALSE	FALSE	FALSE	FALSE		citation	<a href="https://schema.org/distribution">https://schema.org/distribution</a>
	otherId	TRUE	FALSE	FALSE	FALSE		citation	
	otherIdAgency	FALSE	FALSE	FALSE	FALSE	otherId	citation	
	otherIdValue	FALSE	FALSE	FALSE	FALSE	otherId	citation	
	author	TRUE	FALSE	TRUE	TRUE		citation	<a href="http://purl.org/dc/terms/creator">http://purl.org/dc/terms/creator</a>
	authorName	FALSE	TRUE	TRUE	TRUE	author	citation	
	authorAffiliation	FALSE	TRUE	TRUE	FALSE	author	citation	
	authorIdentifierScheme	FALSE	FALSE	TRUE	FALSE	author	citation	<a href="http://purl.org/spar/datacite/AgentIdentifierScheme">http://purl.org/spar/datacite/AgentIdentifierScheme</a>
	authorIdentifier	FALSE	FALSE	TRUE	FALSE	author	citation	<a href="http://purl.org/spar/datacite/AgentIdentifier">http://purl.org/spar/datacite/AgentIdentifier</a>

Source: [Dataverse Metadata Schema](#)

# Metadata Citation block as RDF

```
@prefix citation: <https://dataverse.org/schema/citation/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

citation: citation:accessToSources [ citation:schema_description "Level of documentation of the original sources." ;
    citation:schema_displayOrder 7.7e+01 ;
    citation:schema_fieldType "textbox" ;
    citation:schema_metadatablock_id "citation" ;
    citation:schema_name "accessToSources" ;
    citation:schema_title "Documentation and Access to Sources" ] ;
citation:alternativeTitle [ citation:schema_description "A title by which the work is commonly referred, or an abbreviation of the title." ;
    citation:schema_displayOrder 2e+00 ;
    citation:schema_fieldType "text" ;
    citation:schema_metadatablock_id "citation" ;
    citation:schema_name "alternativeTitle" ;
    citation:schema_title "Alternative Title" ] ;
citation:alternativeURL [ citation:schema_description "A URL where the dataset can be viewed, such as a personal or project website. " ;
    citation:schema_displayFormat "<a href=\"#VALUE\" target=\"_blank\">#VALUE</a>" ;
    citation:schema_displayOrder 3e+00 ;
    citation:schema_fieldType "url" ;
    citation:schema_metadatablock_id "citation" ;
    citation:schema_name "alternativeURL" ;
    citation:schema_title "Alternative URL" ;
    citation:schema_watermark "Enter full URL, starting with http://" ] ;
citation:author [ skos:broader citation:authorAffiliation,
    citation:authorIdentifier,
    citation:authorName ;
    citation:authorAffiliation [ citation:schema_advancedSearchField "True" ;
        citation:schema_description "The organization with which the author is affiliated." ;
        citation:schema_displayFormat "(#VALUE)" ;
        citation:schema_displayOrder 9e+00 ;
        citation:schema_displayoncreate "True" ;
        citation:schema_facetable "True" ;
        citation:schema_fieldType "text" ;
        citation:schema_metadatablock_id "citation" ;
        citation:schema_name "authorAffiliation" ;
        citation:schema_parent "author" ;
        citation:schema_title "Affiliation" ] ;
```

Source: [SEMAF client](#) (CLARIAH.nl project)

# Relationships (predicates) in Dataverse metadata schema

**Every predicate such as `skos:exactMatch` refers to the relationship between the subject (fieldname) and object (termURI):**

```
citation:alternativeTitle skos:exactMatch "http://purl.org/dc/terms/alternative" .
citation:alternativeURL skos:exactMatch "https://schema.org/distribution" .
citation:dateOfDeposit skos:exactMatch "http://purl.org/dc/terms/dateSubmitted" .
citation:kindOfData skos:exactMatch "http://rdf-vocabulary.ddialliance.org/discovery#kindOfData" .
citation:otherReferences skos:exactMatch "http://purl.org/dc/terms/references" .
citation:relatedDatasets skos:exactMatch "http://purl.org/dc/terms/relation" .
citation:subject skos:exactMatch "http://purl.org/dc/terms/subject" .
citation:title skos:exactMatch "http://purl.org/dc/terms/title" .
citation:contributor skos:exactMatch "http://purl.org/dc/terms/contributor" ;
    skos:narrower citation:contributorName .

citation:authorAffiliation skos:altLabel "authorAffiliation" ;
    skos:broader citation:author ;
    skos:prefLabel "Affiliation" .

citation:authorIdentifier skos:altLabel "authorIdentifier" ;
    skos:broader citation:author ;
    skos:exactMatch "http://purl.org/spar/datacite/AgentIdentifier" ;
    skos:prefLabel "Identifier" .
```

# Dataverse metadata export in OAI\_ORE format are triples

dcterms:modified:	"2023-08-17"
dcterms:creator:	"Harvard Dataverse"
@type:	"ore:ResourceMap"
▼ @id:	"https://dataverse.harvard.edu/api/datasets/export?exporter=OAI_ORE&persistentId=https://doi.org/10.7910/DVN/63MIDP"
▼ ore:describes:	
▼ author:	
citation:authorName:	"Ravina, Mark"
citation:authorAffiliation:	"University of Texas at Austin"
authorIdentifierScheme:	"ORCID"
authorIdentifier:	"0000-0002-8726-7618"
▼ citation:keyword:	
citation:keywordValue:	"Japanese geography"
▼ citation:dsDescription:	
▼ citation:dsDescriptionValue:	"File descriptions:\n\nVillage_level_calculations.R – calculates village-level metrics\n\nParcels_sample.txt – random sample of 37,295 parcels from 25,000 unique locations, a random sample of a complete data set of 65,201 unique locations\n\nShikoku_Voronoi_map.R – code to generate Voronoi map "Figure_11_Interactive_map_of_Iyo.html"\n\nShikoku_Voronoi_data.txt – data for Shikoku_Voronoi_map.R\n\nngadm40_JPN_shp – folder os shapefiles for Shikoku_Voronoi_map.R\n\nDomain_Simpson_complete.txt – complete domain-level data for logit calculations, based on all 65,201 locations and 97,553 parcels."
▼ citation:datasetContact:	
citation:datasetContactName:	"Ravina, Mark"
citation:datasetContactAffiliation:	"University of Texas at Austin"
citation:datasetContactEmail:	"mr56267@eid.utexas.edu"
▼ title:	"Replication Data for: Algorithmic Maps and the Political Geography of Early-modern Japan"
dateOfDeposit:	"2023-08-17"
▼ subject:	
0:	"Arts and Humanities"
1:	"Social Sciences"
citation:depositor:	"Ravina, Mark"
@id:	"https://doi.org/10.7910/DVN/63MIDP"

Export Metadata ▼

OAI\_ORE  
DataCite  
OpenAIRE  
Schema.org JSON-LD  
DDI  
Dublin Core  
DDI HTML Codebook  
JSON



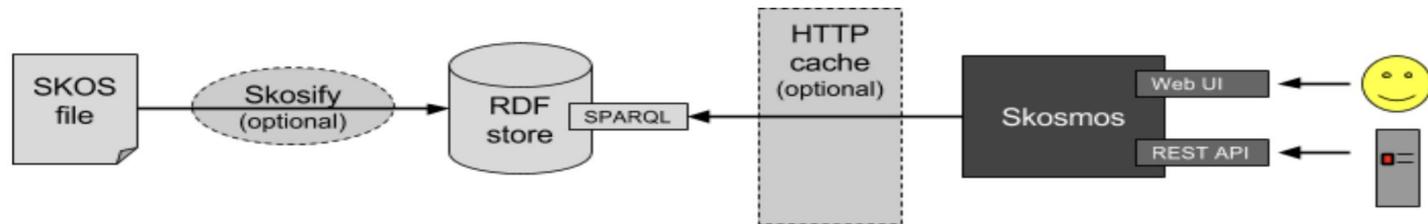
We need more controlled values, ORCID, Wikidata, ...!

Link: [https://dataverse.harvard.edu/api/datasets/export?exporter=OAI\\_ORE&persistentId=doi%3A10.7910/DVN/63MIDP](https://dataverse.harvard.edu/api/datasets/export?exporter=OAI_ORE&persistentId=doi%3A10.7910/DVN/63MIDP)

# External CV support as a key to FAIR interoperability

DANS has developed CV support as plugin solution in the Horizon 2020 funded SSHOC EU project, in the collaboration with GDCC and Harvard IQSS. This integration is based on Skosmos developed by National Library of Finland.

James D. Myers, & Vyacheslav Tykhonov. (2023). A Plug-in Approach to Controlled Vocabulary Support in Dataverse. <https://doi.org/10.5281/zenodo.8133723>



Skosmos architecture

# ELLST thesaurus hosted by Skosmos

## Key features:

- Skosmos is framework powered by **Jena Fuseki** triple store with **SPARQL**
- flexible API with search interface
- export in RDF
- Docker setup
- available as external component in DANS “Archive in a box”
- community support

The screenshot shows the Skosmos interface for the ELSST Thesaurus. The top navigation bar includes 'Vocabularies', 'About', 'Feedback', 'Help', and 'Interface language: English (United States)'. The main content area is titled 'ELSST Thesaurus' and features a search bar with 'Content language' set to 'English'. The left sidebar shows an alphabetical index with 'Hierarchy' selected, and a list of terms including 'ABILITY', 'ABILITY ASSESSMENT', 'ABILITY EVALUATION', 'ABILITY GROUPING', 'ABJURATION OF FAITH', 'ABORIGINAL PEOPLE', 'ABORTION', 'ABSENCE FROM SCHOOL', 'ATTENDANCE', 'ABSENCE FROM WORK', 'ABSENTEEISM (WORK)', 'ABSENT PARENT', 'PARENTAL DEPRIVATION', 'ABSENT PARENTS', 'PARENTAL DEPRIVATION', 'ABSENTEEISM (WORK)', 'ABUSE OF THE ELDERLY', 'ELDER ABUSE', 'ABUSED CHILDREN', 'CHILD ABUSE', 'ACADEMIC ABILITY', 'ACADEMIC ACHIEVEMENT', 'ACADEMIC ADMISSION', 'EDUCATIONAL ADMISSION', 'ACADEMIC APTITUDE', 'ACADEMIC ABILITY', 'ACADEMIC DEPARTMENTS', 'ACADEMIC EVALUATION', 'EDUCATIONAL ASSESSMENT', 'ACADEMIC FACILITIES', 'EDUCATIONAL FACILITIES', 'ACADEMIC FREEDOM', 'ACADEMIC GROUPING', 'EDUCATIONAL GROUPING', 'ACADEMIC RECORDS', 'EDUCATIONAL RECORDS', 'ACADEMIC SUCCESS', 'ACADEMIC ACHIEVEMENT', 'ACADEMIC YEAR', 'EDUCATIONAL YEAR', 'ACCENTS (DIALECT)', 'DIALECTS', 'ACCESS TO COMPUTERS', 'ACCESS TO INFORMATION AND COMMUNICATIONS TECHNOLOGY', and 'ACCESS TO COUNTRYSIDE'. The main content area displays the entry for 'RELIGIOUS EXPERIENCE', including its preferred term, broader concept ('RELIGION'), related concepts ('MEDITATION', 'MYSTICISM'), entry terms ('ABJURATION OF FAITH', 'RELIGIOUS CONVERSION'), identifier ('urn:ddi:int.CESSDA.ELsst:c9d00f39-4285-4911-852f-a2aaf20d3c8b:3'), is a version of ('elsst:c9d00f39-4285-4911-852f-a2aaf20d3c8b (elsst)'), date issued ('2022-09-20'), has previous version ('https://elsst.cessda.eu/id/2/c9d00f39-4285-4911-852f-a2aaf20d3c8b (elsst)'), and in other languages ('NÁBOŽENSKÁ ZKUŠENOST' in Czech, 'RELIGIØS OPLEVELSE' in Danish, 'RELIGIØS OMVENDELSE' in Danish, 'TROSBEKENDELSE' in Danish, 'RELIGIEUZE ERVARING' in Dutch, 'USKONNOLLISET KOKEMUKSET' in Finnish, 'HENGELLINEN HERÄTYS' in Finnish, and 'KÄÄNTYMYS' in Finnish).

Source: [ODISSEI Skosmos](#)

# Dataverse integration with Skosmos is term lookup based

Subject \* ?

Keyword ?

Vocabulary ?

wikidata

Term ?

COVID-19



Vocabulary URL ?

http://www.wikidata.org/entity/Q810689

Vocabulary ?

wikidata

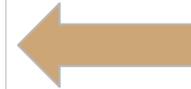
Term ?

covi



Vocabulary URL ?

- COVID-19
- COVID-19 pandemic
- COVID-19 pandemic in India
- Covi
- Covi
- Covi
- SARS-CoV-2



Resulting in TermURI

Related Publication ?

Citation ?

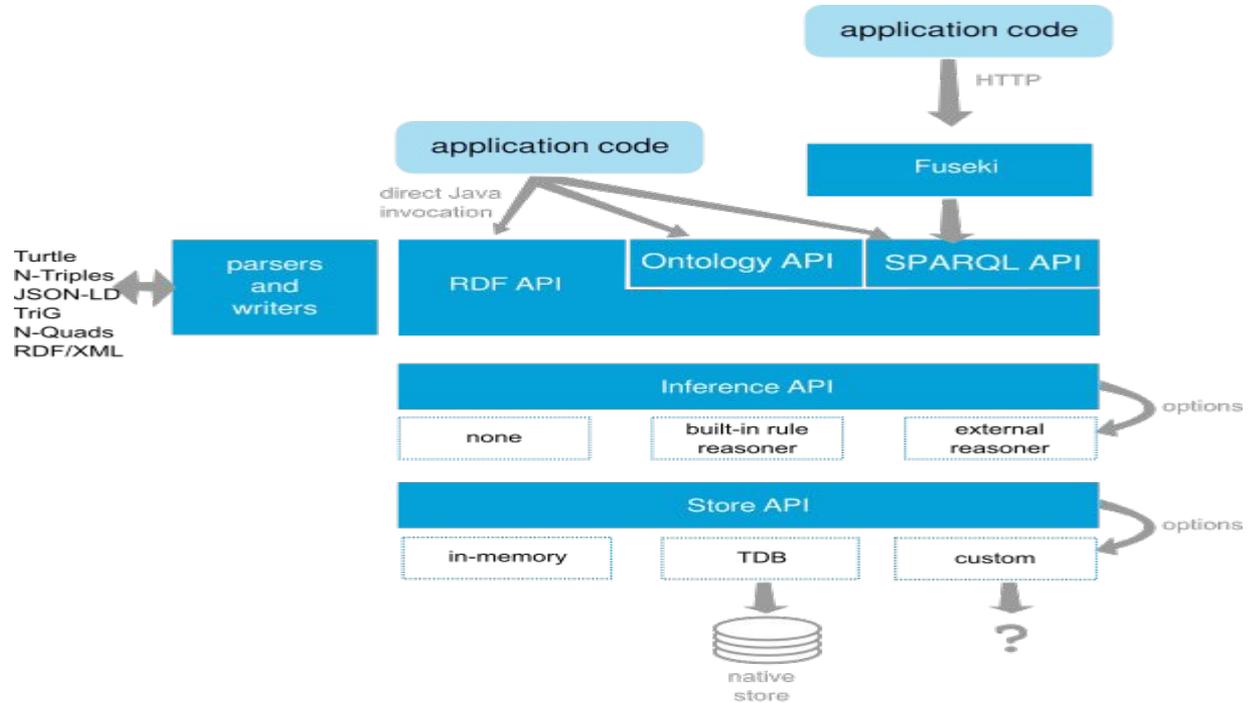


External controlled vocabularies support contributed by [SSHOC project](#) (data infrastructure for the EOSC)

# Meet “built-in” Dataverse triple store: Jena Fuseki

**Jena** stores information as RDF triples in directed graphs, and allows your code to *add, remove, manipulate, store and publish* that information

**Fuseki** is a SPARQL server that provides REST-style API endpoint using the SPARQL protocol over HTTP



# Querying Dataverse as Linked Open Data (ODISSEI project)

## SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

### EXAMPLE QUERIES

Selection of triples

Selection of classes

### PREFIXES

rdf

rdfs

owl

xsd



### SPARQL ENDPOINT

/odissei/query

### CONTENT TYPE (SELECT)

JSON

### CONTENT TYPE (GRAPH)

Turtle

```
5 PREFIX citation: <https://dataverse.org/schema/citation/>
6 PREFIX ore: <http://www.openarchives.org/ore/1.0/terms/>
7 PREFIX distributor: <https://dataverse.org/schema/citation/distributor#>
8 PREFIX variable: <https://portal.staging.odissei.nl/schema/variableInformation/variable#>
9 SELECT ?id ?title ?distributorname
10 WHERE {
11   ?id dcterms:title ?title;
12       dcterms:subject ?subject;
13       schema:datePublished ?date;
14       citation:Distributor ?distributor.
15   ?distributor distributor:Name ?distributorname.
16   ?id schema:datePublished "2010-01-01"
17 }
```

### QUERY RESULTS



Table

Raw Response



Showing 1 to 1 of 1 entries

Search:

Show

50



entries

id	title	distributorname	
1	<doi:10.57934/0b01e41080216906>	"Minimumloonklasse van banen van werknemers"	"Centraal Bureau voor Statistiek"

# Using grlc to manage SPARQL queries through github



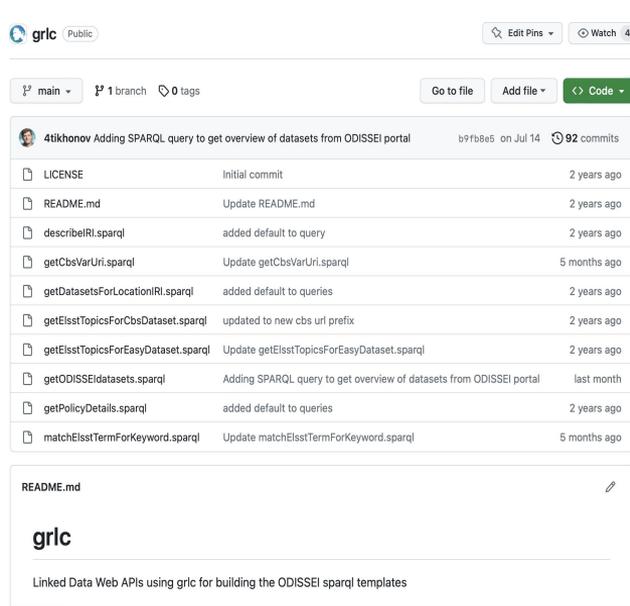
grlc makes all your [Linked Data](#) accessible to the Web by automatically converting your SPARQL queries into RESTful APIs. With (almost) no effort! Simply:

- 1 Create a GitHub repository, and store all your SPARQL queries in there (like in [this example](#)). If you don't have a GitHub account, [go get one](#). You can also just write down the username and the repository name of somebody else :-)
- 2 Go to the address bar of this page, and append `/api/github_username/repository_name` to it. So if I want the API derived from GitHub's username `foo` and repository `bar`, I append `/api/foo/bar/` to the domain name (`/api/foo/bar/api-docs` will work too). Now hit enter. Done!

[Show me an example SPARQL repo](#)

[Show me the equivalent API](#)

Take a look at the increasing number of [users](#) and [SPARQL repositories](#) on GitHub that are using **grlc** to generate APIs!



grlc Public

main 1 branch 0 tags

Go to file Add file Code

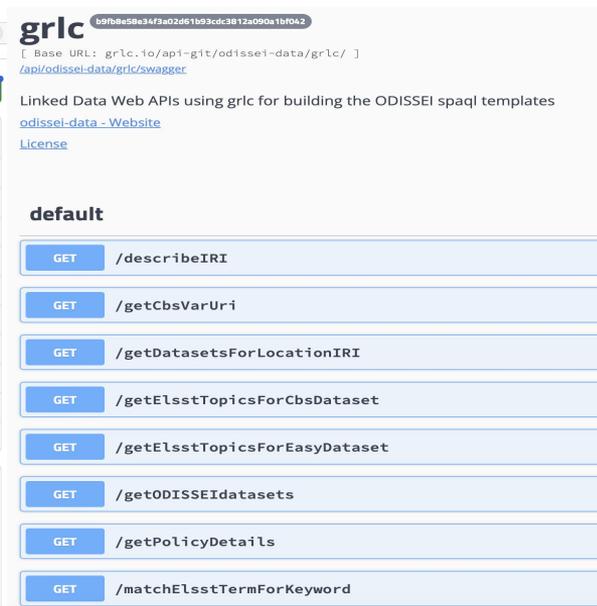
4tkhonov Adding SPARQL query to get overview of datasets from ODISEI portal b9fb8e5 on Jul 14 92 commits

LICENSE	Initial commit	2 years ago
README.md	Update README.md	2 years ago
describeIRI.sparql	added default to query	2 years ago
getCbsVarUri.sparql	Update getCbsVarUri.sparql	5 months ago
getDatasetsForLocationIRI.sparql	added default to queries	2 years ago
getElstTopicsForCbsDataset.sparql	updated to new cbs url prefix	2 years ago
getElstTopicsForEasyDataset.sparql	Update getElstTopicsForEasyDataset.sparql	2 years ago
getODISEIdatasets.sparql	Adding SPARQL query to get overview of datasets from ODISEI portal	last month
getPolicyDetails.sparql	added default to queries	2 years ago
matchElstTermForKeyword.sparql	Update matchElstTermForKeyword.sparql	5 months ago

README.md

## grlc

Linked Data Web APIs using grlc for building the ODISEI sparql templates



grlc b9fb8e58e34f8a02d61b93cdc3b12a090a1bf042

[ Base URL: `grlc.io/api-git/odissei-data/grlc/` ]  
[/api/odissei-data/grlc/swagger](#)

Linked Data Web APIs using grlc for building the ODISEI sparql templates  
[odissei-data - Website](#)  
[License](#)

### default

- GET `/describeIRI`
- GET `/getCbsVarUri`
- GET `/getDatasetsForLocationIRI`
- GET `/getElstTopicsForCbsDataset`
- GET `/getElstTopicsForEasyDataset`
- GET `/getODISEIdatasets`
- GET `/getPolicyDetails`
- GET `/matchElstTermForKeyword`

Link: [https://grlc.io/api/odissei-data/grlc#/default/get\\_getODISEIdatasets](https://grlc.io/api/odissei-data/grlc#/default/get_getODISEIdatasets)

# grlc usage example in ODISSEI

SPARQL query stored in github:

GET /getODISSEIdatasets

```
#+ endpoint: http://fuseki.experimental.odissei.nl/odissei/sparql

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <https://schema.org/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX citation: <https://dataverse.org/schema/citation/>
PREFIX ore: <http://www.openarchives.org/ore/1.0/terms/>
SELECT ?id ?title ?description ?keyword
WHERE {
  ?id dcterms:title ?title;
      citation:Description ?descid;
      citation:Keyword ?keyid.
  ?descid ?descstype ?description.
  ?keyid ?k ?keyword
}
```

Response from SPARQL transformed by grlc in specified format:

Request URL

```
https://grlc.io/api-git/odissei-data/grlc/getODISSEIdatasets?endpoint=http%3A%2F%2Ffuseki.experimental.odissei.nl%2Fodissei%2Fsparql
```

Server response

Code Details

200

Response body

```
id,title,description,keyword
doi:10.17026/dans-xdu-8j3m,Green Xenophobia Project,<p>The experimental study is conducted to investigate the public opinion on how immigration leads to overpopulation and in turn might damage the sustainable development and environmental protection.</p>,Demography and population
```

# Dataverse metadata semantic enrichment with Skosmos CVs

The screenshot shows the Dataverse ODISSEI Portal interface. At the top, the Dataverse logo and navigation links (Search, User Guide, Support, Sign Up, Log In) are visible. The main header includes the CBS (Centraal Bureau voor de Statistiek) logo and the text 'ODISSEI Portal >'. A search bar contains the term 'education' and a 'Q' icon, with a link to 'Advanced Search'. Below the search bar, there are filters for 'Dataverses (0)' and 'Datasets (105)'. A 'Publication Year' filter shows results for 2021 (1), 2020 (1), 2019 (3), 2018 (5), and 2017 (6). A 'Keyword Term' filter shows 'Onderwijs (16)' and several related terms. The search results section displays '1 to 10 of 105 Results' and a 'Sort' dropdown. Two results are shown: 1. 'Adult Education Survey- Onderzoek onder volwassenen naar het volgen van opleidingen (lange, korte)' by Centraal Bureau voor de Statistiek, 2011. The description states the goal is to provide information on adult education participation. The variable is 'EDUCATION @en'. 2. 'Scholieren en studenten in een kalenderjaar' by Centraal Bureau voor de Statistiek, 1999. The description is a file of pupils and students attending education in a calendar year. The variable is 'EDUCATION @en'.

Multilingual (semantic) search: keyword 'education' isn't included in dataset metadata!

# Semantic Search for the Dutch data landscape (ODISSEI project)

In addition to social sciences metadata from CBS, LISS, DataverseNL and DANS the metadata from the Historical Sample of the Netherlands (HSN) is now included. The Portal also harvests the newly released [DANS Data Station Social Sciences and Humanities](#) which has replaced the DANS archive EASY.

## Improved search for CBS metadata

In this updated version of the ODISSEI Portal prototype, improvements are made to better find metadata related to CBS microdata sets. There is much more information included on the variables which are now visible in the metadata block "Variable information". It is now also possible to search for metadata records based on the short names used in the CBS Remote Access environment (e.g. KINDOUDERTAB).

## Multilingual search for CBS keywords

A great improvement is that the Portal now supports multilingual search for CBS keywords. The CBS metadata is enriched by mapping the available Dutch keywords to keywords in the European Social Sciences Language Thesaurus (ELSST). In ELSST translations are available into other languages and this allows you to find Dutch metadata records when searching with English terms. For example, when you are in the CBS part of the ODISSEI Portal, you can search for "employment" and find datasets related to this, even though the metadata is only available in Dutch. This feature will be extended and improved in the next few months.

## Further developments

Until the end of 2024, the Portal will be further developed and improved by ODISSEI, DANS, VU and SURF as part of the ODISSEI Roadmap project. The collected metadata will be enriched with existing controlled vocabularies to increase the findability of the information. More metadata providers will be added to the Portal with the ultimate goal of giving researchers access to information about all relevant social science datasets in the Netherlands. Data providers from the ODISSEI community are encouraged to get in touch with the Portal team to make their metadata available in the Portal. Last but not least, the functionalities around the data access broker will be developed, which should enable users to request data directly from the various data providers.

The screenshot displays the ODISSEI Portal search interface. At the top, there is a navigation bar with the 'Dataverse' logo, a search input field, and links for 'User Guide', 'Support', and 'Log In'. Below the navigation bar, the 'CBS (CBS)' logo is visible, along with a link to the 'ODISSEI Portal'. The main search area features a search input field containing the text 'companies' and a search button. To the right of the search input, there is a link to 'Advanced Search'. Below the search input, there are filters for 'Dataverses (0)' and 'Datasets (173)'. The search results are displayed in a list format, showing the first 10 of 173 results. The first result is titled 'Milieukosten van bedrijven' and is dated Jan 1, 2012. The second result is titled 'Milieukosten van bedrijven, publicatiebestand' and is also dated Jan 1, 2012. The third result is titled 'Bedrijfseenheden en de koppelingen naar de ondernemingsgroep' and is dated Jan 1, 2000. Each result includes a brief description and a 'Keyword Term'.

ODISSEI portal updated with Multilingual search engine: <https://dans.knaw.nl/en/news/improved-search-and-more-data-to-find/>

# KG ecosystem: LOD Summarizer as “prompt” for SPARQL

## did-summarizer

Linked Data summarizer driven by Decentralized Identifiers (DIDs)

Developed by DANS Labs, funded by [CLARIAH project](#).

The main aim of the Summarizer service is to gain an overview about which vocabularies are already used in CLARIAH, or might be useful in CLARIAH. The core is to build an analyzing pipeline containing data collection, vocabulary analysis, report. While concentrating on the automatised pipeline, we also indicate at which point where expert/manual curation is needed.

To bridge between different knowledge domains it is needed to find communalities, cross-walks, mappings between vocabularies. A prerequisite for this is to gain an overview what vocabularies exist (VOCABULARY part) and how they are used (DATASET part). Despite of many existing registries this problem is by no means solved, nor are there standard, off-the shelf solutions for gaining such an overview.

Decentralized identifiers (DIDs) are being used to create resolvable globally accessible unique and persistent identifiers to support various Linked Data tasks in FAIR way:

- assign DID to SPARQL query to make it persistent
- cache vocabulary concept content and relationships
- store and get statistics of usage for vocabulary concepts available in the time dimension, just like the Internet Archive
- assign unique DID to other services used in CLARIAH pipelines ...

<https://github.com/Dans-labs/did-summarizer>

```
curl -X 'GET' \  
'http://0.0.0.0:8001/summarizer?url=https%3A%2F%2Fraw.githubusercontent.com%2FAKSW%2Fdssn.rdf%2  
-H 'accept: application/json'
```

Response body:

```
{  
  "statements": {  
    "statements": 119,  
    "unique objects": 64,  
    "unique predicates": 25,  
    "unique subjects": 22  
  },  
  "prefixes": {  
    "http://www.w3.org/2000/01/rdf-schema#": "rdfs",  
    "http://www.w3.org/2002/07/owl#": "owl",  
    "http://www.w3.org/2004/02/skos/core#": "skos",  
    "http://xmlns.com/foaf/0.1/": "foaf",  
    "http://rdfs.org/sioc/ns#": "sioc",  
    "http://usefulinc.com/ns/doap#": "doap",  
    "http://www.w3.org/2003/06/sw-vocab-status/ns#": "vs",  
    "http://purl.org/dc/terms/": "dct",  
    "http://purl.org/net/dssn/": "dssn"  
  },  
  "stats": {  
    "dssn": 126,  
    "rdfs": 60,  
    "vs": 13,  
    "doap": 1,  
    "foaf": 20,  
    "owl": 15,  
    "dct": 1,  
    "sioc": 11,  
    "skos": 1  
  }  
}
```

# KG ecosystem: controlled vocabulary recommender

Vocabulary Recommender Command-line interface (CLI) was developed by [Triply](#) and provides a recommendation interface which returns relevant Internationalized Resource Identifiers (IRIs) based on the search input. It works with SPARQL or Elasticsearch endpoints which contain relevant vocabulary datasets.

[DANS](#) has created API service out of it.

Usage example:

```
curl -X 'GET' \  
'http://0.0.0.0:8001/recommend?searchTerm=person&searchClass=class' \  
-H 'accept: application/json'
```

Response:

```
[  
  [  
    {  
      "searchTerm": "person",  
      "vocabs": [  
        "https://w3id.org/pnv#"  
      ],  
      "homogeneous": [  
        {  
          "iri": "https://w3id.org/pnv#Person",  
          "score": 1,  
          "vocabPrefix": "https://w3id.org/pnv#",  
          "vocabDomain": "https://w3id.org/pnv#",  
          "description": "A Person is a human being whose individual existence can somehow be doc",  
          "category": "class"  
        }  
      ],  
      "single": [  
        {  
          "iri": "http://xmlns.com/foaf/0.1/Person",  
          "score": 0.8853529652138054,  
          "vocabPrefix": "foaf",  
          "vocabDomain": "http://xmlns.com/foaf/0.1/",  
          "description": "A person.",  
          "category": "class"  
        }  
      ],  
    }  
  ],  
]
```

# Knowledge Graphs as Supervisor for LLM

*“You’ll see the real power of the usage of Large Language Models when you’ll be able to “talk” to data just like to your assistant, or even ask questions to your own documents”.*

Anton Polishko, ex-Google, CoronaWhy

# LLaMA: A foundational, 65-billion-parameter large language model

Querying on historical facts is OK

```
~ Tue Aug 22 2023 01:28 am
ollama run stable-beluga:70b
>>> Who was the 45th president in the US? Give response as RDF triples
  Subject: 45th President of the United States
  Predicate: is_a
  Object: Donald Trump

>>> Now list other 5 last presidents. Also as RDF triples
  Subject: 44th President of the United States
  Predicate: is_a
  Object: Barack Obama

  Subject: 43rd President of the United States
  Predicate: is_a
  Object: George W. Bush

  Subject: 42nd President of the United States
  Predicate: is_a
  Object: Bill Clinton

  Subject: 41st President of the United States
  Predicate: is_a
  Object: George H.W. Bush

  Subject: 40th President of the United States
```

Does know nothing about current events

```
curl http://172.28.112.1:8081/v1/chat/completions \
-H "Content-Type: application/json" \
-d '{
  "messages": [
    { "role": "user", "content": "### Who is winner of Eurovision in 2023? \n### Response:" }
  ],
  "temperature": 0.5,
  "max_tokens": -1,
  "stream": false
}'
{
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "message": {
        "content": " I'm sorry, but the Eurovision Song Contest for 2023 has not taken place yet. The winner will be determined in the future when the competition occurs.",
        "role": "assistant"
      }
    }
  ],
  "created": 1692705863,
  "id": "chatcmpl",
  "model": "LLaMA_CPP",
  "object": "chat.completion",
  "truncated": false,
  "usage": {
    "completion_tokens": 38,
    "prompt_tokens": 52,
    "total_tokens": 90
  }
}
```

YouTube: <https://www.youtube.com/watch?v=m-phwvbbgXE>

# MuseIT collections powered by Now.Museum

The screenshot displays the Dataverse web interface. At the top, the Dataverse logo is on the left, and navigation links for Search, User Guide, Support, Sign Up, and Log In are on the right. Below the header, a message states: "Test collection prepared by the MuseIT project using the Now.Museum technology." A search bar contains the text "Search this dataverse..." and a "Q" icon, with a link to "Advanced Search". On the left sidebar, there are filters for "Dataverses (0)", "Datasets (3,718)", and "Files (0)". Below these are filters for "Publication Year" (2023 (3,718)), "Author Name" (MuseIT (3,718)), "Subject" (Arts and Humanities (3,718)), and "Keyword Term" (media (1,954), song contest (863), eurovision song (784), Eurovision (562), grand final (473)). A "More..." link is at the bottom of the sidebar. The main content area shows "1 to 10 of 3,718 Results" and a "Sort" dropdown. Three search results are visible, each with a document icon, a title, a date, a snippet, and a DOI link. The first result is titled "esc\_charts: Eurovision Charts@esc\_charts-31mMost-streamed #Eurovision 2023 songs on..." and dated Aug 22, 2023. The second result is titled "Eurovision winner shows the power of music for nursing home residents" and dated Aug 21, 2023. The third result is titled "Duncan Laurence: Ο νικήτης της Eurovision 2019 παντρεύτηκε με τον επί χρόνια σύντροφό του" and dated Aug 21, 2023.

- MuseIT is Horizon 2020 project to co-design and develop an inclusive multisensory platform with interactive technologies for people with disabilities to have enriched engagement with cultural assets and experiences.
- Now.Museum is Dataverse based hosting platform for MuseIT, CoronaWhy and other projects. In the collaboration with Yves Rozenholc, University Paris Cite.
- The technology powered by shared knowledge graph

# Event/News registration in Now.Museum Dataverse

## Eurovision 2023

(MuselT)

Root > MuselT > Eurovision 2023 >

## Loreen calls Eurovision win 'surreal'

Version 1.0



MuselT, 2023, "Loreen calls Eurovision win 'surreal'", <https://doi.org/10.5072/FK2/JWI30C>, Root, V1

[Cite Dataset](#) ▾

[Learn about Data Citation Standards.](#)

Contact Owner

Share

Dataset Metrics 

0 Downloads 

### Description

Loreen calls Eurovision win 'surreal'Eurovision winner Loreen has said she is still comprehending her 'surreal' victory'. The Swedish popstar claimed her second victory in Liverpool on Saturday with song Tattoo. She became the first woman to win the contest twice, having first won in 2012 with Euphoria.Popular VideosWatch MoreMore VideosHeadlinesSunday World NewsletterSign up for the latest news and updatesMore VideosDownload the Sunday World appNow download the free app for all the latest Sunday World News, Crime, Irish Showbiz and Sport. Available on Apple and Android devicesLatest News

### Subject

Arts and Humanities

### Keyword

loreen calls, calls eurovision, eurovision win, surreal'eurovision winner, winner loreen, swedish popstar, popstar claimed, second victory, song tattoo, contest twice, euphoria.popular videoswatch, videoswatch moremore, moremore videosheadlinesunday, videosheadlinesunday world, world newslettersign, latest news, updatesmore videosdownload, sunday world, world appnow, appnow download, free app, latest sunday, world news, irish showbiz, android deviceslatest, deviceslatest news, media

### License/Data Use Agreement

[Custom Dataset Terms](#)

# Assistant prompt in LLaMA

Prompt

```
curl http://172.28.112.1:8081/v1/chat/completions \  
-H "Content-Type: application/json" \  
-d '{  
  "messages": [  
    { "role": "user", "content": "### Instruction: Give me search terms helping to find out who was the Eurovision winner in 2023? Do not mention names of winners. \n### Give response in JSON format." }  
  ],  
  "temperature": 0.5,  
  "max_tokens": -1,  
  "stream": false  
'
```



Response

```
{  
  "choices": [  
    {  
      "finish_reason": "stop",  
      "index": 0,  
      "message": {  
        "content": " {\n      \"searchTerms\": [\n        \"Eurovision winner\",\n        \"2023\",\n        \"winner\"\n      ]\n}",  
        "role": "assistant"  
      }  
    }  
  ],  
  "created": 1692705207,  
  "id": "chatcmpl",  
  "model": "LLaMA_CPP",  
  "object": "chat.completion",  
  "truncated": false,  
  "usage": {  
    "completion_tokens": 37,  
    "prompt_tokens": 75,  
    "total_tokens": 112  
  }  
}
```

# SPARQL query to find out who won Eurovision in 2023

PREFIXES

rdf rdfs owl xsd +

SPARQL ENDPOINT: /nowmuseum/sparql

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```
5 PREFIX citation: <https://dataverse.org/schema/citation/>
6 PREFIX ore: <http://www.openarchives.org/ore/1.0/terms/>
7 PREFIX id: <https://doi.org/>
8 SELECT distinct ?title
9 WHERE {
10   ?id schema:distribution ?url;
11       dcterms:title ?title;
12       citation:dsDescription ?description;
13       citation:keyword ?keywordID.
14   ?description citation:dsDescriptionValue ?summary.
15   ?keywordID citation:keywordValue ?keyword.
16   filter contains(?title,"Eurovision")
17   filter contains(?keyword,"winner")
18   filter contains(?title,"2023")

```

QUERY RESULTS

Table Raw Response

Showing 1 to 1 of 1 entries

Search:  Show 100 entries

	title
1	"Sweden wins Eurovision Song Contest 2023 with Tattoo by Loreen"

Showing 1 to 1 of 1 entries

# Interested? Contact me!

Slava Tykhonov, R&D  
(DANS-KNAW, the Netherlands)  
[vyacheslav.tykhonov@dans.knaw.nl](mailto:vyacheslav.tykhonov@dans.knaw.nl)