Data Archiving and Networked Services

DANS

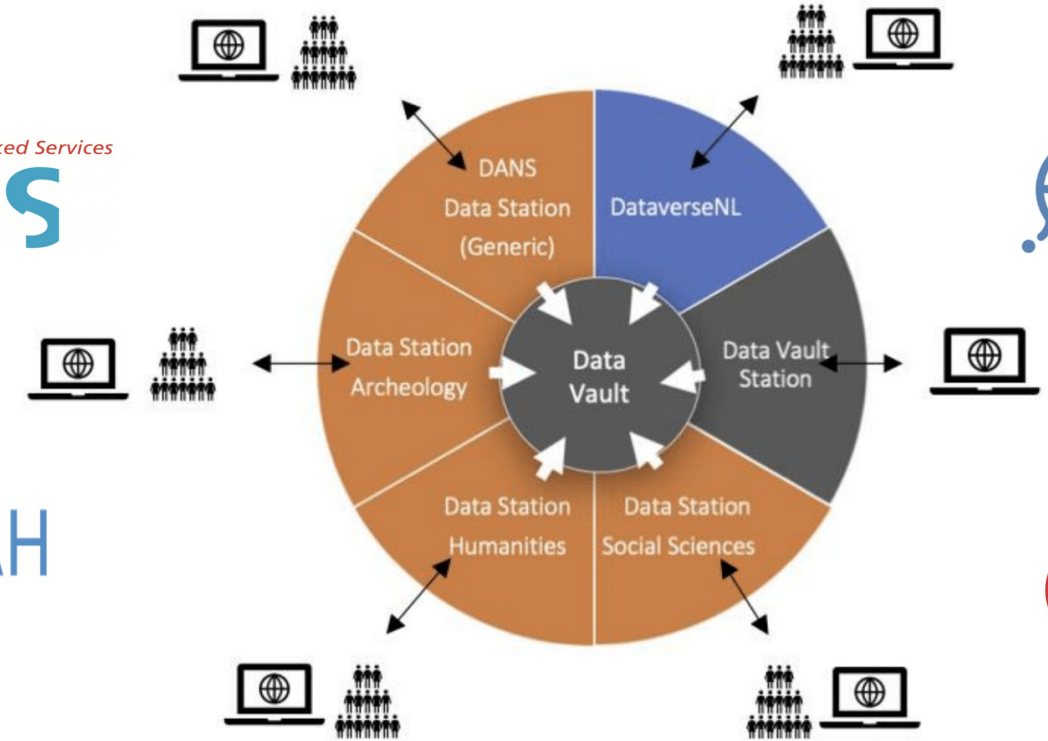
Knowledge Graphs and Semantic Search in Dataverse

Dataverse Community call, Harvard IQSS

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

22 August 2023

DANS Data Stations - Future Data Services



Data Archiving and Networked Services
DANS

SSHOC
social sciences & humanities open cloud

CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

EOSC
SYNERGY

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

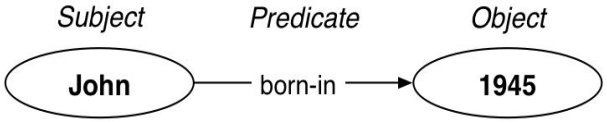
What is knowledge graph and how it can help?

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:[11]

- *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.[12]
- *General structure*: A network of entities, their semantic types, properties, and relationships.[13][14]
- *Supporting reasoning over inferred ontologies*: A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.[2]

Source: Wikipedia

Semantic triple and Resource Description Framework (RDF)

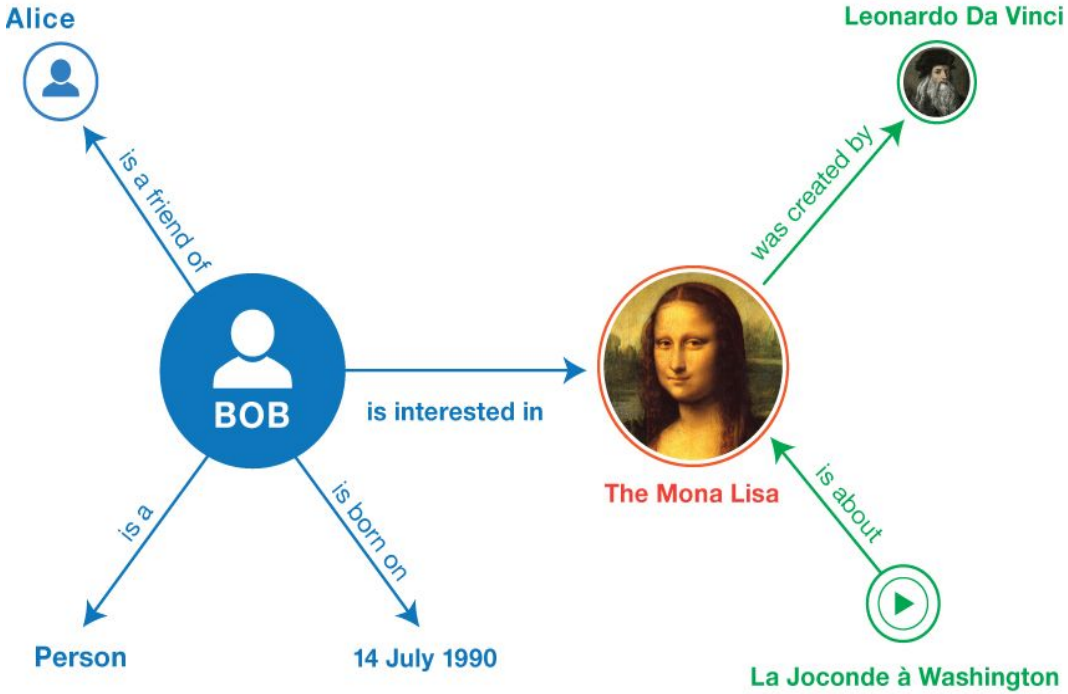


“This format enables knowledge to be represented in a machine-readable way. Particularly, every part of an RDF triple is individually addressable via unique URIs — for example, the statement "Bob knows John" might be represented in RDF as:

```
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	http://purl.org/dc/terms/title
	subtitle	FALSE	FALSE	FALSE	FALSE		citation	
	alternativeTitle	FALSE	FALSE	FALSE	FALSE		citation	http://purl.org/dc/terms/alternative
	alternativeURL	FALSE	FALSE	FALSE	FALSE		citation	https://schema.org/distribution
	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	http://purl.org/dc/terms/creator
	authorName	FALSE	TRUE	TRUE	TRUE	author	citation	
	authorAffiliation	FALSE	TRUE	TRUE	FALSE	author	citation	
	authorIdentifierSche	FALSE	FALSE	TRUE	FALSE	author	citation	http://purl.org/spar/datacite/AgentIdentifierScheme
	authorIdentifier	FALSE	FALSE	TRUE	FALSE	author	citation	http://purl.org/spar/datacite/AgentIdentifier

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" ] ;
```

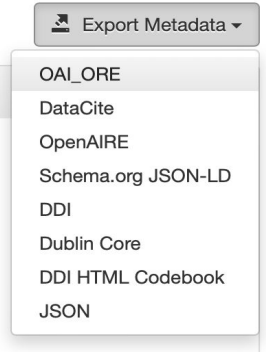
Source: SEMAF client (CLARIAH.nl project)

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\ngadm40_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"
```

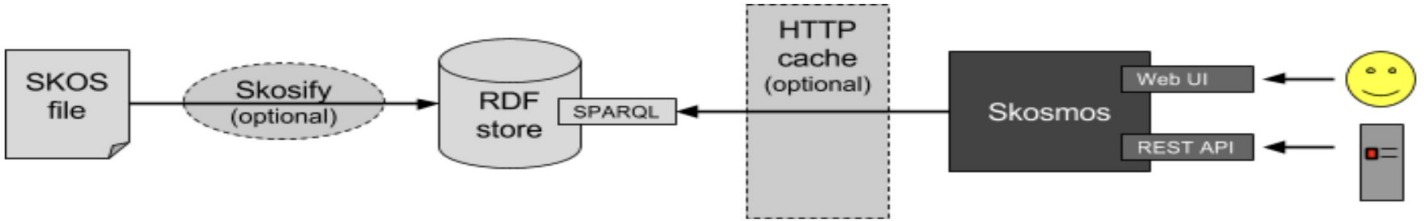

We need more controlled values, ORCID, Wikidata, …!

Link: 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

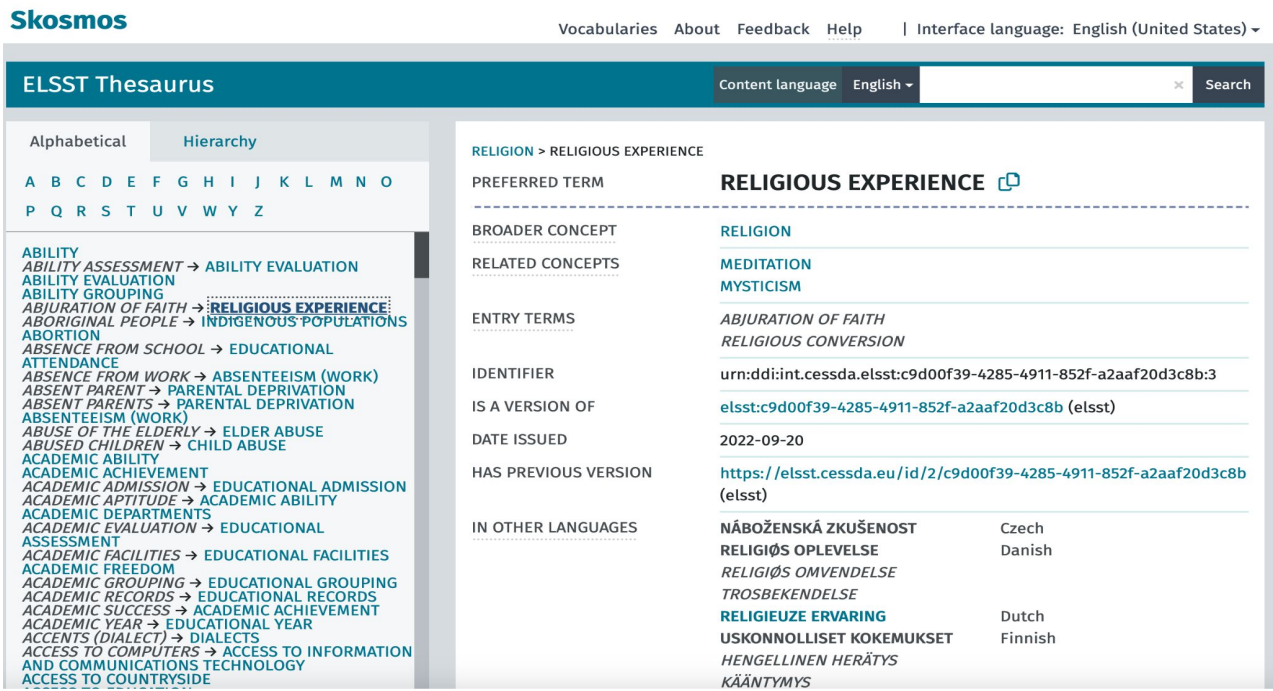


Skosmos architecture

ELLST thesaurus hosted by Skosmos

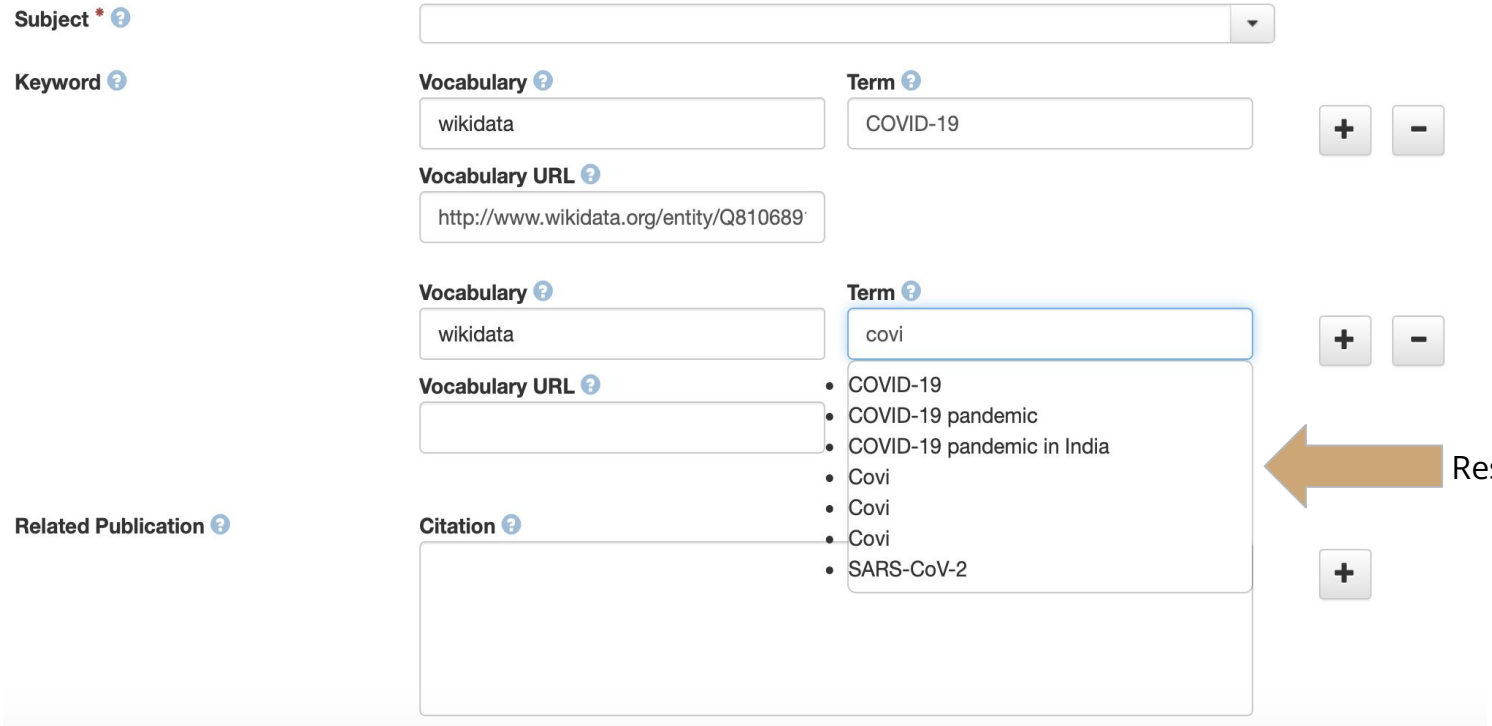
Key features:

- 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



Source: [ODISSEI Skosmos]

Dataverse integration with Skosmos is term lookup based



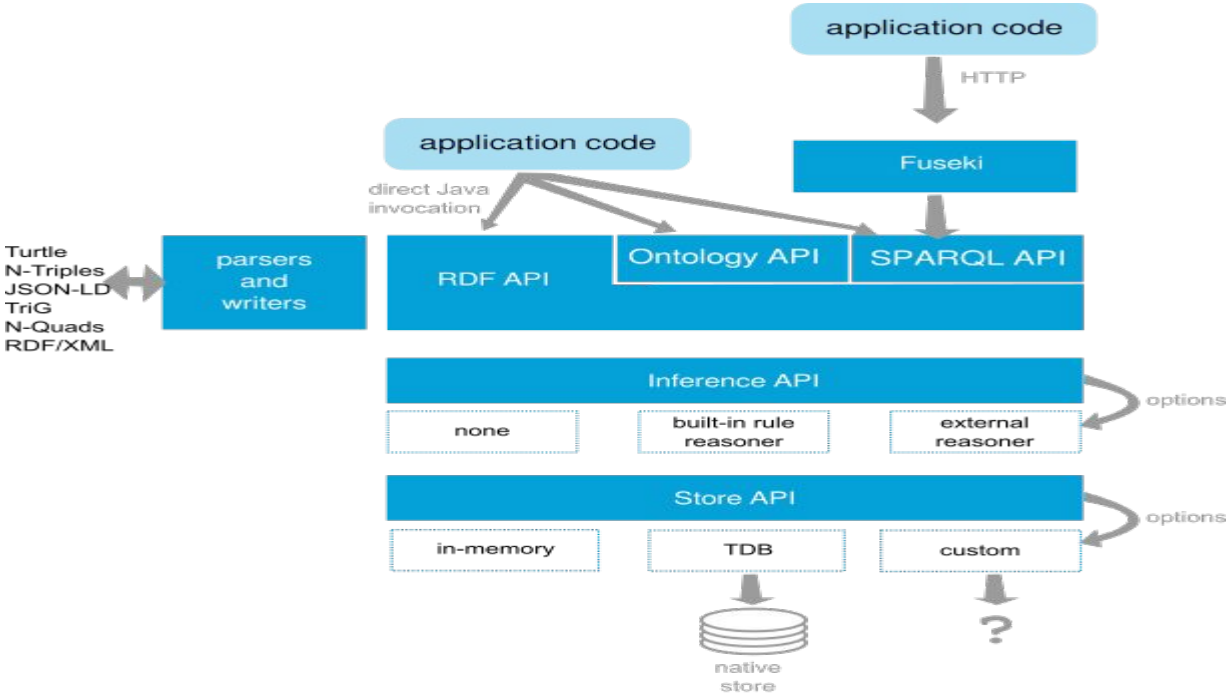
Resulting in TermURI

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

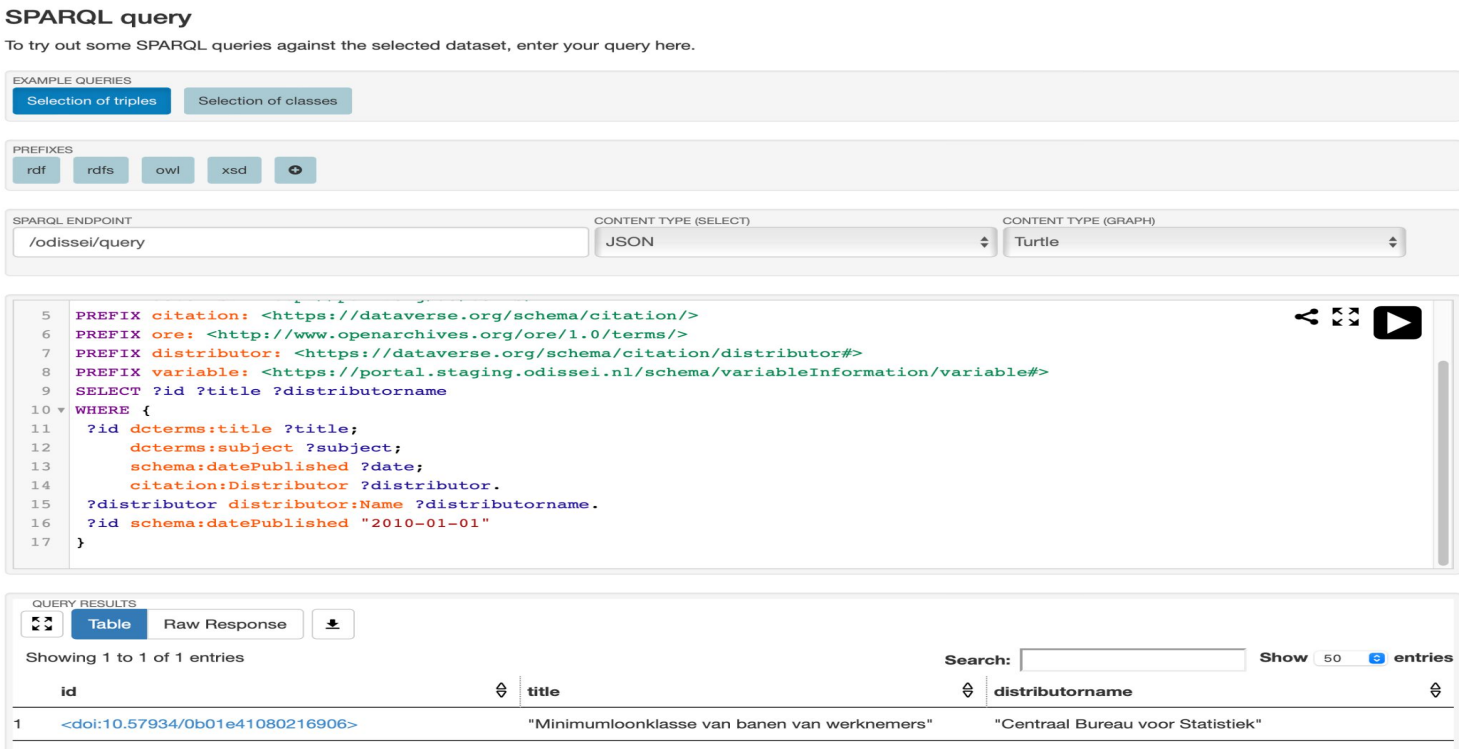
Meet “built-in” Dataverse triple store: Jena Fuseki

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

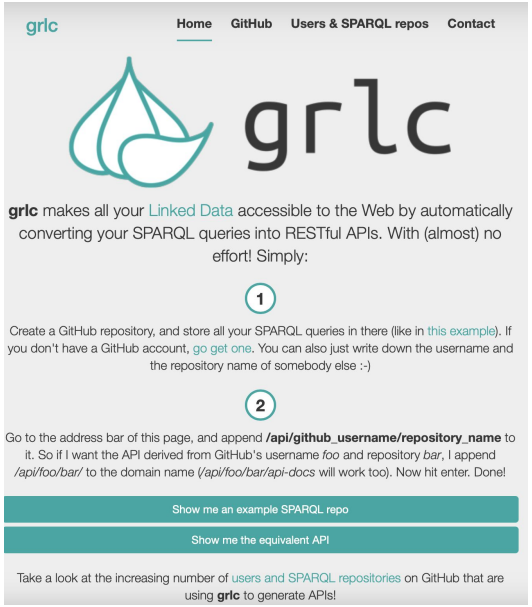
Fuseki is a SPARQL server that provides REST-style API endpoint using the SPARQL protocol over HTTP

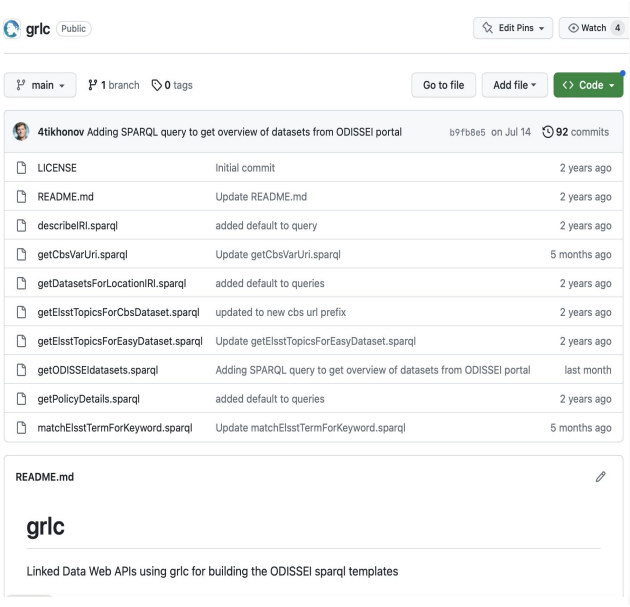


Querying Dataverse as Linked Open Data (ODISSEI project)



Using grlc to manage SPARQL queries through github





Link: https://grlc.io/api/odissei-data/grlc#/default/get_getODISSEIdatasets

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 ?desctype ?description.
 ?keyid ?k ?keyword
}
```

Response from SPARQL transformed by grlc in specified format:

Request URL

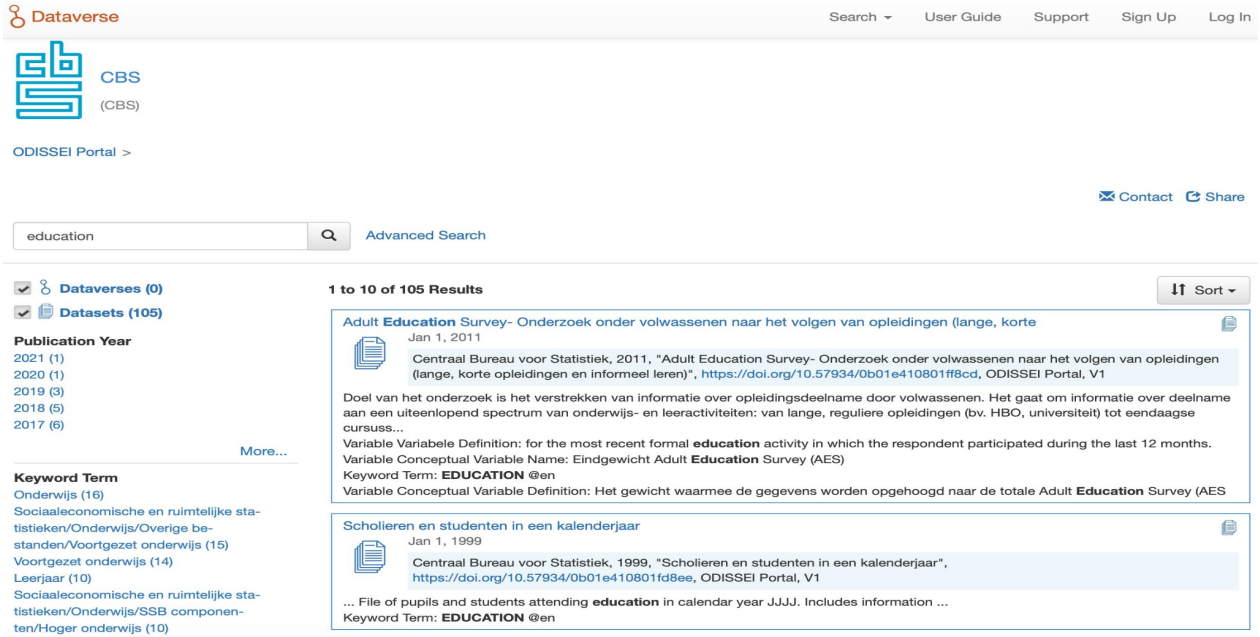
```
https://grlc.io/api-git/odissei-data/grlc/getODISSEIdatasets?endpoint=http%3A%2F%2Ffuseki.experimental.odissei.nl%2Fodissei%2Fsparql
```

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



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

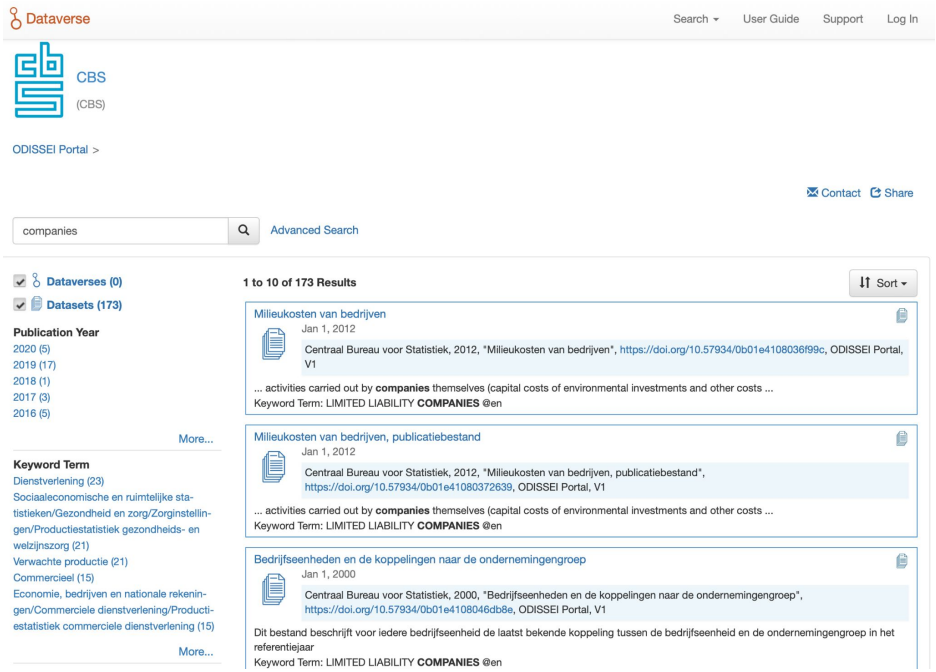
In this updated version of the ODISSEI Portal prototype, improvements are made to better find metadata related to CBS microdata sets. There is much more information included on the variables which are now visible in the metadata block "Variable information". It is now also possible to search for metadata 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.



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

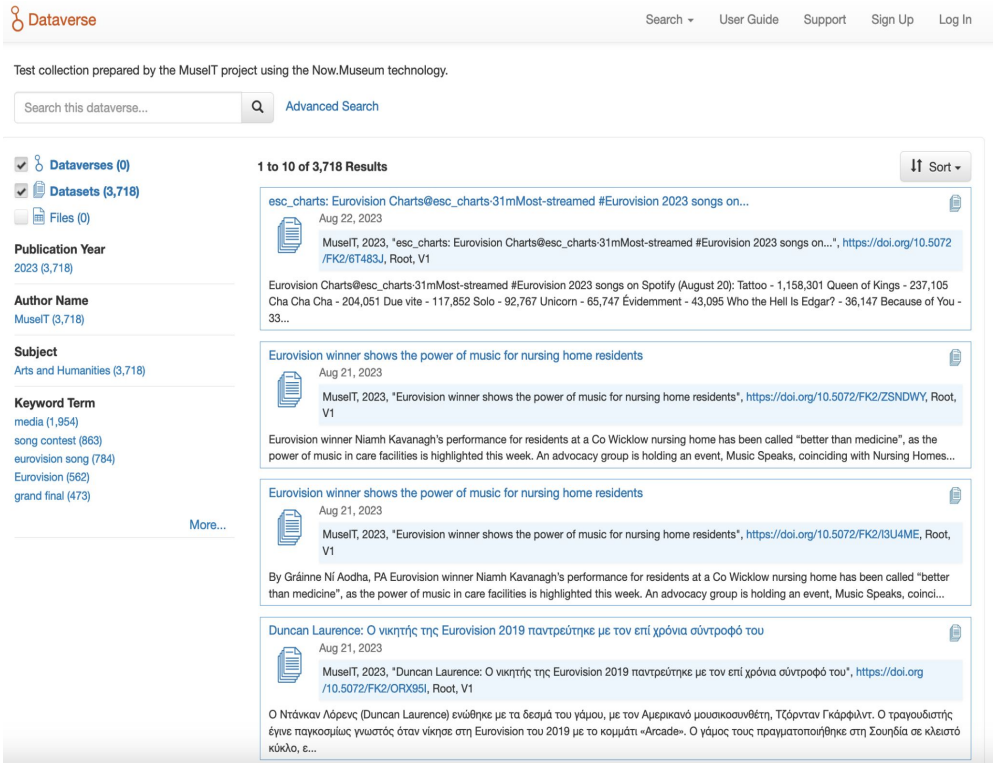
```
~ Tue Aug 22 2023 01:28 am
ollama run stable-beluga:70b
>>>  Who was the 45th president in the US? Give response as RDF triples
 Subject: 45th President of the United States
Predicate: is_a
Object: Donald Trump

>>> 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
  }
}
```

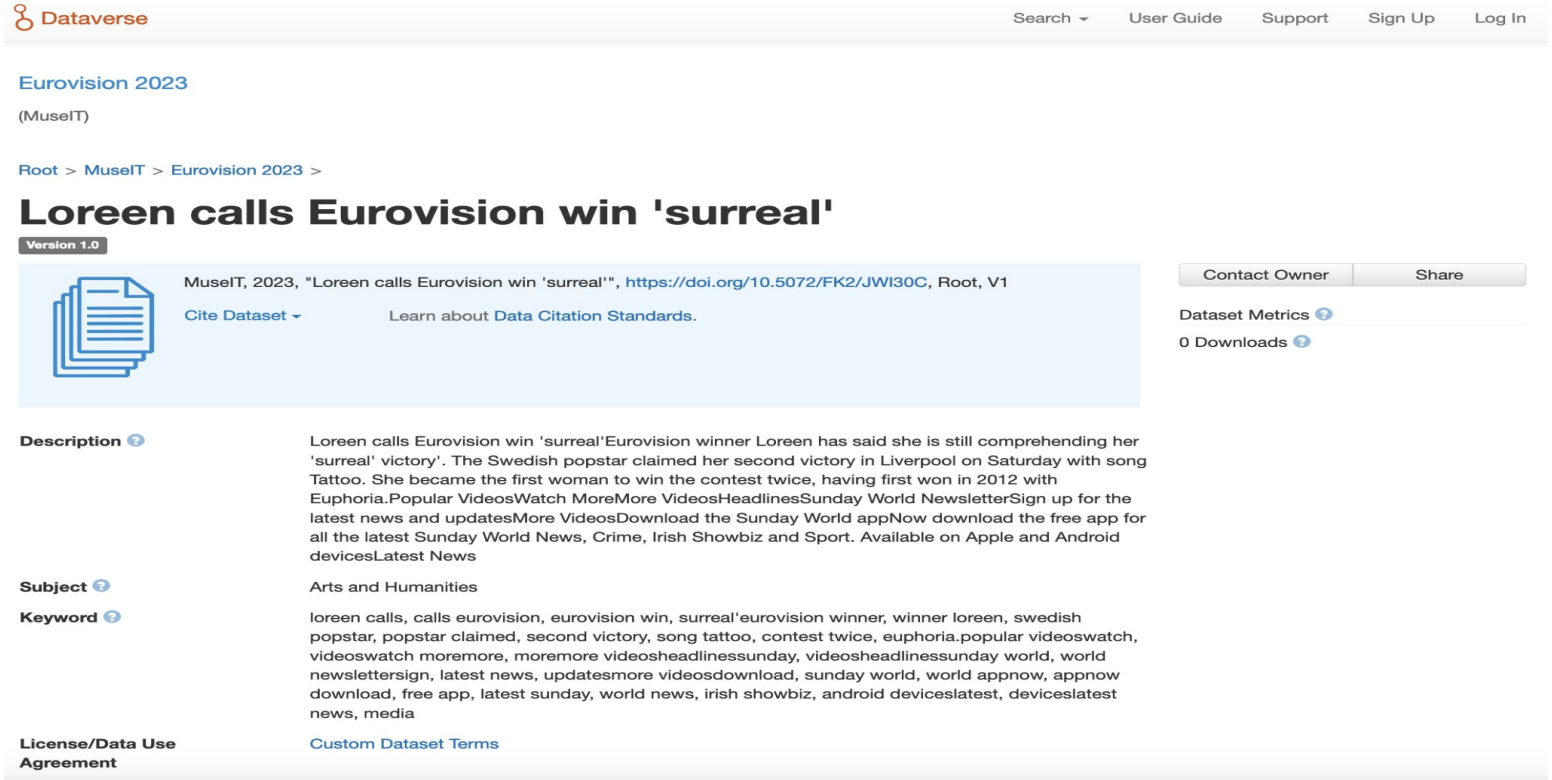
YouTube: https://www.youtube.com/watch?v=m-phwvbbgXE

MuseIT collections powered by Now.Museum



- 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



Assistant prompt in LLaMA

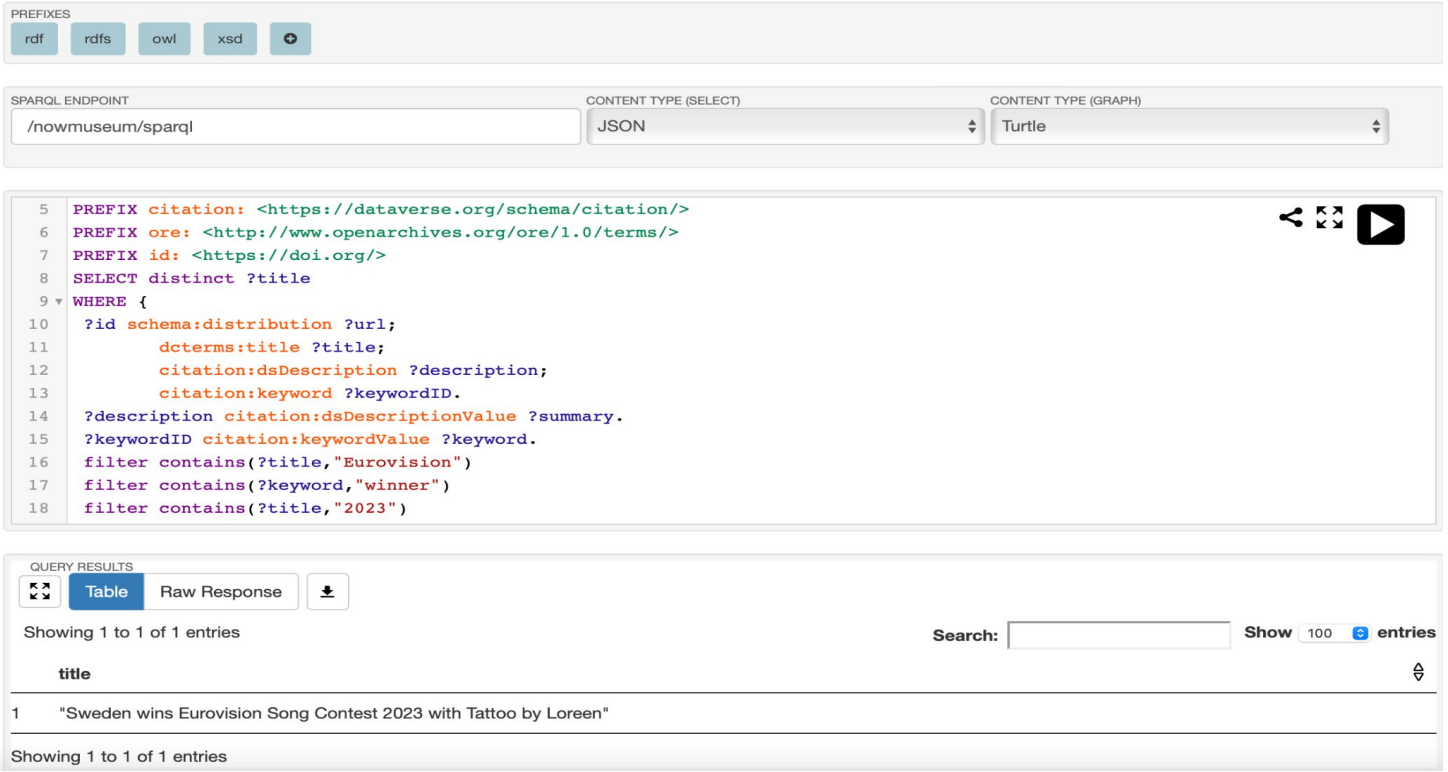
Prompt

```
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 winner
s. \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



Interested? Contact me!

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