



**SSHOC**  
social sciences & humanities open cloud

## External controlled vocabularies support in Dataverse

Slava Tykhonov (DANS-KNAW)  
lead software engineer, DANS R&D

**Dataverse Community Meeting 2021**

**16 June 2021**

Harvard University



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Project:



# SSHOC

social sciences & humanities open cloud



Horizon 2020  
European Union Funding  
for Research & Innovation

**Type of action & funding:**  
Research and Innovation action  
(INFRAEOSC-04-2018)

**Partners: 47**

(20 beneficiaries + 27 LTPs)

SSH ESFRI Landmarks and Projects  
& international SSH data infrastructures

**Project budget:**

€ 14,455,594.08

**Duration: 40 months**

(January 2019 – 30 April 2022)

**Project website:**  
[www.SSHOpenCloud.eu](http://www.SSHOpenCloud.eu)

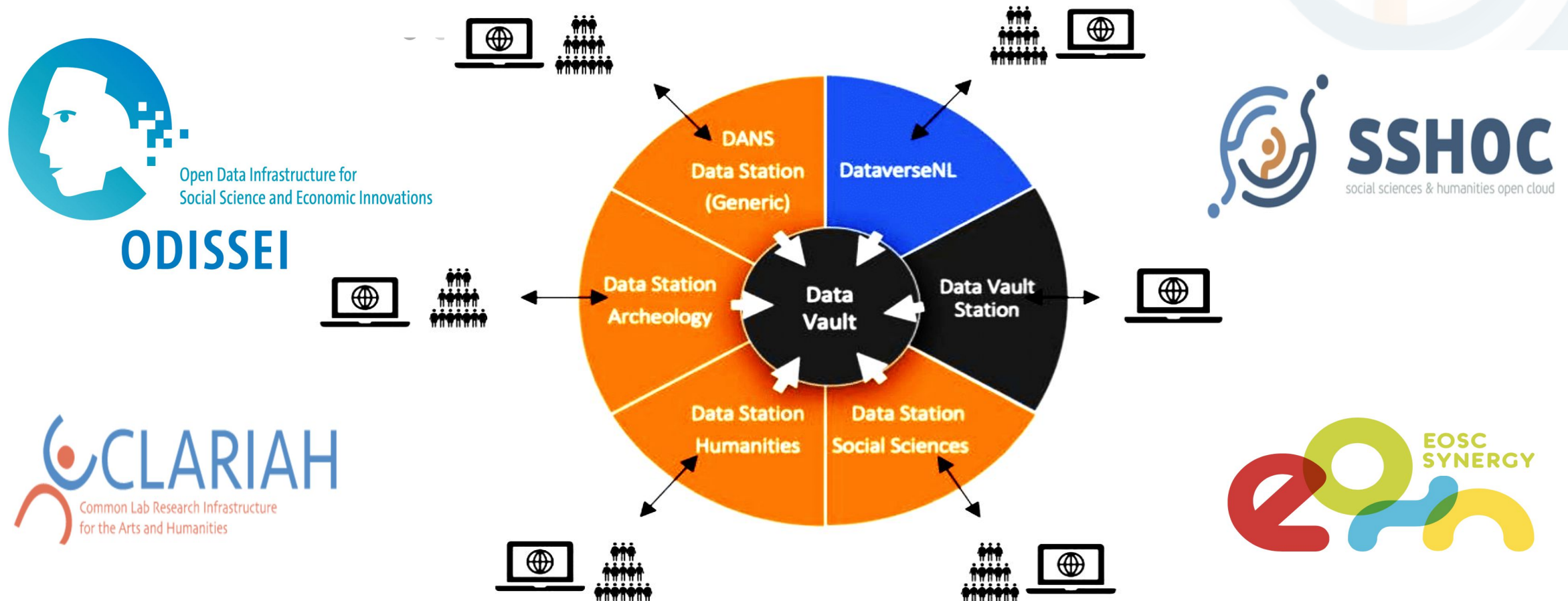


## Objectives:

- creating the social sciences and humanities (SSH) part of European Open Science Cloud (EOSC)
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC



# DANS Data Stations - Future Data Services



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

# FAIR and Dataverse

FM [AID*]	Question	Dataverse Q'aire	Dataverse Optimized
Identifier type	1	DOI	DOI
F1A	2		
F1B	Not tested in Q'aire		
F2A	4A		
F2A	4B		
F3	5B		
F4	6A		
F4	6B		
A1.1	7A		
A1.2	8A		
A1.2	8B	N/A	N/A
A2	9		
I1	10		
I2	11		
I3	12		
R1.1	13		
R1.2	14A		

## DATAVERSE FAIR SUMMARY

- **Strong support for Findable, Accessible, and Reusable principles**
- **Weak for Interoperable principles**
- In agreement\* with FAIR test results (\*F3 was fixed after test)
- There is no FAIR “compliance”
- Instead, it’s a process and can always be improved

Source:

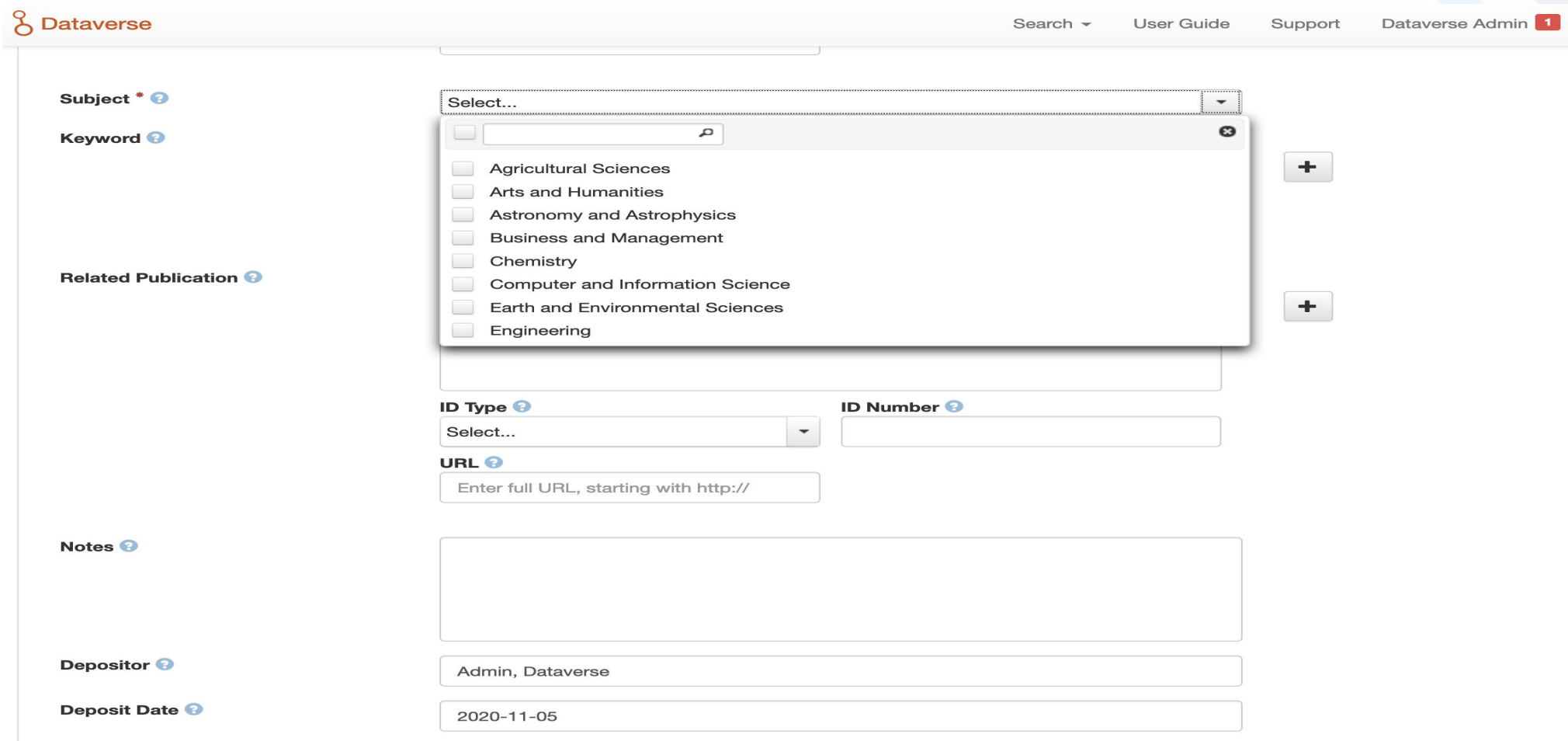
[Mercè Crosas,](#)  
[“FAIR principles and beyond: implementation in Dataverse”](#)

# Out of the box CV support in Dataverse (1)

1	#metadataBlock	name	dataverseAlias	displayName	blockURI	
2		citation		Citation Metadata	<a href="https://dataverse.org">https://dataverse.org</a>	
3	#datasetField	name	title	description	watermark	fieldType
82	#controlledVocabularies	DatasetField	Value	identifier	displayOrder	
83		subject	Agricultural Sciences	D01	0	
84		subject	Arts and Humanities	D0	1	
85		subject	Astronomy and Astrophysics	D1	2	
86		subject	Business and Management	D2	3	
87		subject	Chemistry	D3	4	
88		subject	Computer and Informatics	D7	5	
89		subject	Earth and Environmental Sciences	D4	6	
90		subject	Engineering	D5	7	
91		subject	Law	D8	8	
92		subject	Mathematical Sciences	D9	9	
93		subject	Medicine, Health and Life Sciences	D6	10	
94		subject	Physics	D10	11	
95		subject	Social Sciences	D11	12	
96		subject	Other	D12	13	
97		publicationIDType	ark		0	

Source: [Dataverse Metadata Schema](#)

# Out of the box CV support in Dataverse (2)



The screenshot shows the Dataverse metadata form with a dropdown menu open for the 'Subject' field. The dropdown lists the following categories:

- Agricultural Sciences
- Arts and Humanities
- Astronomy and Astrophysics
- Business and Management
- Chemistry
- Computer and Information Science
- Earth and Environmental Sciences
- Engineering

Other fields visible in the form include:

- Keyword**: A search input field.
- Related Publication**: A field with a plus sign (+) to add related publications.
- ID Type**: A dropdown menu with 'Select...' as the current value.
- ID Number**: A text input field.
- URL**: A text input field with the placeholder 'Enter full URL, starting with http://'.
- Notes**: A large text area for additional information.
- Depositor**: A text input field containing 'Admin, Dataverse'.
- Deposit Date**: A text input field containing '2020-11-05'.

Internal vocabularies are stored in Dataverse, we need more CVs!

# The importance of standards and ontologies

Generic controlled vocabularies to link metadata in the bibliographic collections are well known: ORCID, GRID, GeoNames, Getty.

Medical knowledge graphs powered by:

- Biological Expression Language (BEL)
- Medical Subject Headings (MeSH®) by U.S. National Library of Medicine (NIH)
- Wikidata (Open ontology) - Wikipedia

Integration based on metadata standards:

- MARC21, Dublin Core (DC), Data Documentation Initiative (DDI)

The most of prominent ontologies already available as a Web Services with API endpoints.

# 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”.





# Global Research Identifier Database (GRID) in SKOS

```
<http://www.grid.ac/institutes/grid.1001.0> a skos:Concept ;
  rdfs:label "Australian National University"@en ;
  isni:id "0000 0001 2180 7477"@en ;
  dc:date "1946-01-01"@en ;
  dcterms:identifier "grid.1001.0"@en ;
  vivo:abbreviation "ANU"@en ;
  skos:code "grid.420434.5" ;
  skos:exactMatch "http://www.wikidata.org/entity/Q127990" ;
  skos:inScheme "http://www.grid.ac/schema#CS000" ;
  skos:memberOf cw:CO007 ;
  skos:prefLabel "Australian National University"@en ;
  vcard:Address "http://www.grid.ac/institutes/grid.1001.0/address-0" ;
  foaf:homepage "http://www.anu.edu.au/"@en .

<http://www.grid.ac/institutes/grid.1002.3> a skos:Concept ;
  rdfs:label "Monash University"@en ;
  isni:id "0000 0004 1936 7857"@en ;
  dc:date "1958-01-01"@en ;
  dcterms:identifier "grid.1002.3"@en ;
  skos:code "grid.420434.5" ;
  skos:exactMatch "http://www.wikidata.org/entity/Q598841" ;
  skos:inScheme "http://www.grid.ac/schema#CS000" ;
  skos:memberOf cw:CO007 ;
  skos:prefLabel "Monash University"@en ;
  vcard:Address "http://www.grid.ac/institutes/grid.1002.3/address-0" ;
  foaf:homepage "http://www.monash.edu/"@en .

<http://www.grid.ac/institutes/grid.10025.36> a skos:Concept ;
  rdfs:label "University of Liverpool"@en ;
  isni:id "0000 0004 1936 8470"@en ;
  dc:date "1882-01-01"@en ;
  dcterms:identifier "grid.10025.36"@en ;
  skos:code "grid.420434.5" ;
  skos:exactMatch "http://www.wikidata.org/entity/Q499510" ;
  skos:inScheme "http://www.grid.ac/schema#CS000" ;
  skos:memberOf cw:CO007 ;
  skos:prefLabel "University of Liverpool"@en ;
  vcard:Address "http://www.grid.ac/institutes/grid.10025.36/address-0" ;
  foaf:homepage "http://www.liv.ac.uk/"@en .
```

We already have a lot of data in the global Dataverse network.

Can we provide **depositors** a convenient web interface to link their metadata to external controlled vocabularies?

Is it possible to disambiguate concepts and create links automatically?

# SKOSMOS framework to discover ontologies

Skosmos

Vocabularies About Feedback Help | in English

Global Research Identifier Database GRID


Content language English  Search

Alphabetical Hierarchy Groups

A Á Â Ã Ä Å B C Č Ç D E É F G H  
I Í J K L Ł M N O Ó Ö Ø Õ P Q  
R Ř S Ś Š Ş T U Ú Û V W X Y Z  
Ž H !\* 0-9

Dassault Aviation (France)  
Dassault Systèmes (Canada)  
Dassault Systèmes (France)  
Dassault Systèmes (Germany)  
Dassault Systèmes (Japan)  
Dassault Systèmes (United Kingdom)  
Dassault Systèmes (United States)  
Dat (Norway)  
Data & Society Research Institute  
Data Access Technologies → Model Driven Solutions (United States)  
**Data Archiving and Networked Services**  
Data Assurance and Communication Security  
Data Assurance and Communication Security Research Center → Data Assurance and Communication Security  
Data Fusion International (Ireland)  
Data Fusion Research Center  
Data Harbor (United States)  
Data Management (Italy)  
Data Management Services (United States)  
Data Numerica Institute (United States)  
Data Observation Network for Earth → DataONE  
Data One Global (United States)  
Data Power Decisions (United States)  
Data Respons (Norway)  
Data Sciences International (United States)  
Data Security Council of India  
Data Storage Institute  
Data Voice Exchange (United States)  
Data-Mate (Finland)  
DATA4 (France)  
Data61  
Data:Lab Munich (Germany)  
Databank (Italy)  
Datachassi (Sweden)  
DataCite  
Datacon (Czechia)  
Datacorp (United States)

**Data Archiving and Networked Services**

PREFERRED TERM 

ENTRY TERMS *Netherlands Institute for Permanent Access to Digital Research Resources*


DATE 2005-01-01

IDENTIFIER grid.500519.8

HOMEPAGE <https://dans.knaw.nl/en>

EXACTLY MATCHING <http://www.wikidata.org/entity/Q13570995>

CONCEPTS

URI <http://www.grid.ac/institutes/grid.500519.8> 

Download this concept: [RDF/XML](#) [TURTLE](#) [JSON-LD](#)

- SKOSMOS is developed in Europe by the National Library of Finland (NLF)
- active global user community
- search and browsing interface for SKOS concept
- multilingual vocabularies support
- used for different use cases (publish vocabularies, build discovery systems, vocabulary visualization)

# SKOSMOS API specification in Swagger



## 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

GET	<code>/vocabulary/</code>	General information about the vocabulary
GET	<code>/vocabulary/types</code>	Information about the types (classes) of objects in the vocabulary
GET	<code>/vocabulary/topConcepts</code>	Top concepts of the vocabulary
GET	<code>/vocabulary/data</code>	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.
GET	<code>/vocabulary/search</code>	Finds concepts and collections from a vocabulary by query term
GET	<code>/vocabulary/lookup</code>	Look up concepts by label
GET	<code>/vocabulary/vocabularyStatistics</code>	Number of Concepts and Collections in the vocabulary
GET	<code>/vocabulary/labelStatistics</code>	Number of labels by language
GET	<code>/vocabulary/index/</code>	Initial letters of the alphabetical index

Source: [Finto API](#)

# SKOSMOS API example for GRID ontology

```
{
  ▼ "@context": {
    skos: http://www.w3.org/2004/02/skos/core#,
    isothes: http://purl.org/iso25964/skos-thes#,
    onki: http://schema.onki.fi/onki#,
    uri: "@id",
    type: "@type",
    ▼ results: {
      "@id": "onki:results",
      "@container": "@list"
    },
    prefLabel: "skos:prefLabel",
    altLabel: "skos:altLabel",
    hiddenLabel: "skos:hiddenLabel"
  },
  uri: "",
  ▼ results: [
    ▼ {
      uri: http://www.grid.ac/institutes/grid.500519.8,
      ▼ type: [
        "skos:Concept",
        "foaf:Organization",
        http://www.grid.ac/ontology/Facility
      ],
      localname: "grid.500519.8",
      prefLabel: "Data Archiving and Networked Services",
      lang: "en",
      altLabel: "DANS-KNAW",
      vocab: "grid"
    }
  ]
}
```

# Semantic Gateway as plugin app (in development)

## 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 |

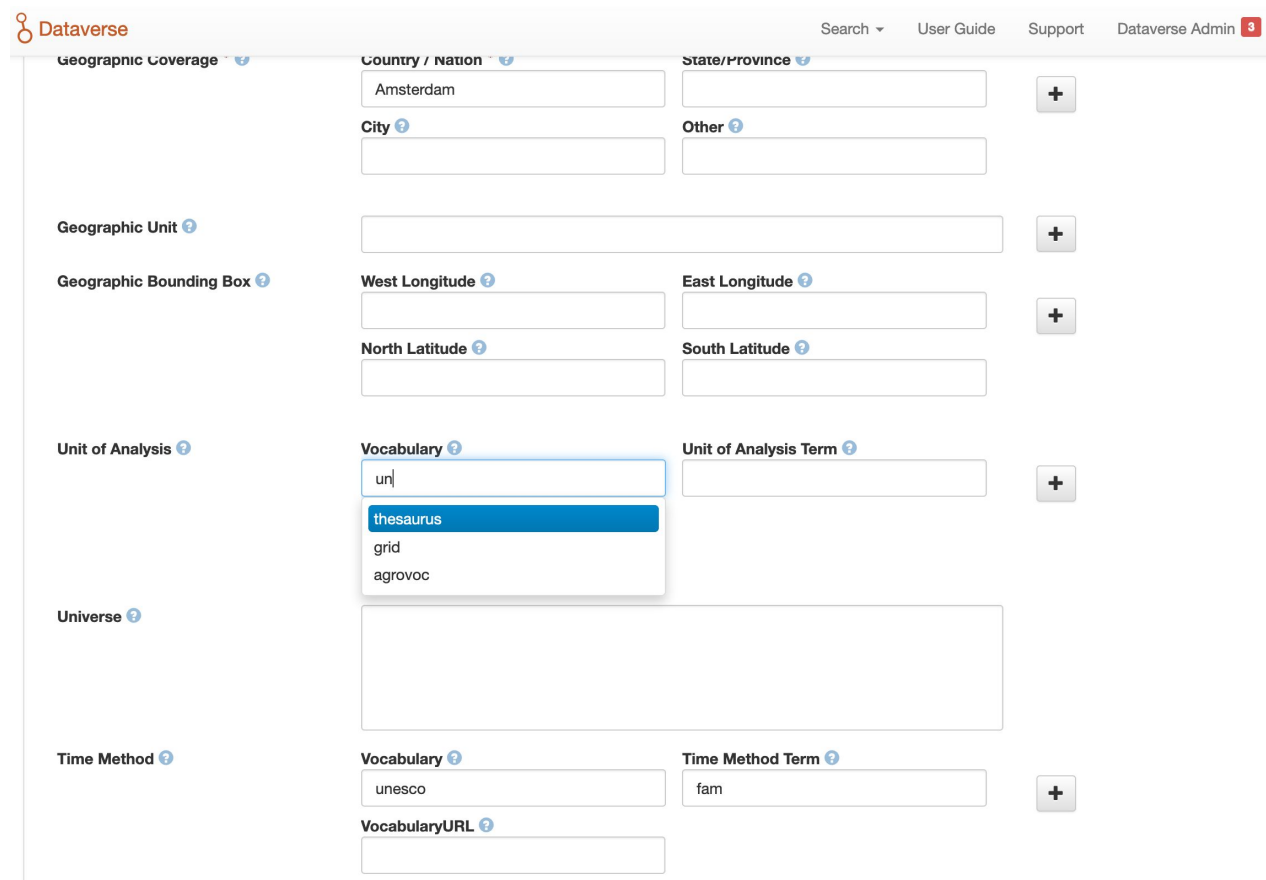
Gateway URL

Dataverse URL

unlock-key

Source: [Dataverse gateway](#)

# Dataverse deposit form with selected CVs



The screenshot displays the Dataverse deposit form interface. The form is organized into several sections, each with a title and a question mark icon for help. The sections and their fields are:

- Geographic Coverage:** Includes fields for Country / Nation (with 'Amsterdam' entered), State/Province, City, and Other. A plus sign (+) is visible to the right.
- Geographic Unit:** A single text input field with a plus sign (+) to the right.
- Geographic Bounding Box:** Includes fields for West Longitude, East Longitude, North Latitude, and South Latitude. A plus sign (+) is visible to the right.
- Unit of Analysis:** Includes a Vocabulary dropdown menu (with 'un|' entered and a list of options: 'thesaurus', 'grid', 'agrovoc') and a Unit of Analysis Term field. A plus sign (+) is visible to the right.
- Universe:** A large empty text area.
- Time Method:** Includes fields for Vocabulary (with 'unesco' entered), Time Method Term (with 'fam' entered), and VocabularyURL. A plus sign (+) is visible to the right.

The Dataverse logo is in the top left corner. The top right corner contains navigation links: Search, User Guide, Support, and Dataverse Admin (with a red notification badge showing '3').

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

# One metadata field linked to many ontologies

The screenshot shows the Dataverse metadata editor interface. At the top, there is a navigation bar with the Dataverse logo, a search dropdown, and links for 'User Guide', 'Support', and 'Dataverse Admin'. The main content area is titled 'Subject \*' and contains a dropdown menu with 'Select...' as the current selection. Below this, there are several rows of metadata fields. Each row consists of a 'Vocabulary' field, a 'Vocabulary URL' field, and a 'Term' field. The 'Vocabulary' and 'Vocabulary URL' fields are text inputs, while the 'Term' field is a text input with a dropdown menu. The 'Term' field is currently open, showing a list of suggestions: 'family', 'family planning', and 'famine'. The 'Term' field is highlighted with a blue border, and the suggestions are listed below it. The 'Vocabulary' and 'Vocabulary URL' fields are also highlighted with a blue border. The 'Term' field is currently empty, and the suggestions are listed below it. The 'Vocabulary' and 'Vocabulary URL' fields are text inputs, while the 'Term' field is a text input with a dropdown menu. The 'Term' field is currently open, showing a list of suggestions: 'family', 'family planning', and 'famine'. The 'Term' field is highlighted with a blue border, and the suggestions are listed below it.

Vocabulary	Vocabulary URL	Term
unesco	http://skos.um.es/unescothes/C01489	Family
thesaurus	http://vocabularyes.irstea.fr/thesaurus/T+	family labour
agrovoc	http://aims.fao.org/aos/agrovoc/c_2785	families
iptc		fam

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

# Improved metadata schema with 4 child fields

cvocDemo Metadata ^

cvocDemo ?

Vocabulary ?  
unesco

Term ?  
social|  
Social adaptation  
Social alienation  
Social and economic rights  
Social and human sciences

Vocabulary URL ?  
http://skos.um.es/unescothes/CS000 +

Term URL ?

Files

**Vocabulary** and **Term** selected by user, **Vocabulary URL** and **Term URL** filled automatically:

cvocDemo Metadata ^

cvocDemo ?

Vocabulary ?  
unesco

Term ?  
Social and economic rights

Vocabulary URL ?  
http://skos.um.es/unescothes/CS000 +

Term URL ?  
http://skos.um.es/unescothes/C03681



# Configuration for external controlled vocabularies

Pull Request to Dataverse core <https://github.com/IQSS/dataverse/pull/7712>

```
[
  {
    "vocab-name": "cvocDemo",
    "minChars": 1,
    "cvoc-url": "https://skosmos.dev.finto.fi/",
    "language": "en",
    "js-url": "/resources/js/cvoc-interface.js",
    "protocol": "skosmos",
    "vocab-uri" : "http://skos.um.es/unescothes/CS000",
    "term-parent-uri": "",
    "vocabs": ["unesco", "stw", "agrovoc"],
    "vocab-codes": ["cvocDemoVocabulary", "cvocDemoTerm", "cvocDemoTermURI", "cvocDemoVocabularyURI"]
  }
]
```

# Javascript interface

```
// depends on jquery ajax
var mapquery = 'prefLabel';
var mapid = 'uri';
var mapping = { query: mapquery, id: mapid };

// autocomplete calls this
function autointerface(request, response, cv, mapping) {
  var protocol = cv.protocol;
  if (!protocol) { protocol = 'skosmos'; } //default
  if (protocol == 'skosmos') {
    return(skosmos(request, response, cv, mapping)); }
  if (protocol == 'example') {
    return(autoexample(request, response)); }
};

function skosmos(request, response, cv, mapping) {
  var termParentUri = "";
  if (cv.termParentUri != "")
    termParentUri = "&parent=" + cv.termParentUri;
  var result = [];
  var tmp = $.ajax({
    url: cv.cvocUrl + '/rest/v1/search?unique=true&vocab=' + cv.selectedVocab + termParentUri,
    dataType: "json",
    data: { query: request.term + '*' },
    success: function(data) {
      var results = data.results;
      var queries = [];
      var array = [];
      $.each(results, function(i, item) {
        queries.push(item.prefLabel);
        array.push({
          value: item[mapping.query],
          id: item[mapping.id]
        });
      });
      response( array );
      console.log( array );
    }
  });
};
```

CV interface implemented as Javascript and placed outside of Dataverse application.

internal:

“js-url”: “/resources/js/cvoc-interface.js”

External:

“js-url”:

“https://raw.githubusercontent.com/Dans-labs/semantic-gateway/main/static/js/interface.js”

# SKOSMOS python module (SKOSMOS-Client)

```
from skosmos_client import SkosmosClient
```

```
# then you can create your own client
```

```
skosmos = SkosmosClient(api_base=http://api.finto.fi/rest/v1/)
```

Finding the available vocabularies:

Vocabulary id: afo                      title: AFO - Natural resource and environment  
ontology

Vocabulary id: allars                    title: Allärs - General thesaurus in Swedish

Vocabulary id: cn                        title: Finnish Corporate Names

Vocabulary id: ic                        title: Iconclass

...

## Other SKOSMOS supported services

- [Finto](#) (Finnish thesaurus and ontology service)
- [CESSDA CV Service](#) has implemented SKOSMOS interface
- [CESSDA ELSST](#) (European Language Social Science Thesaurus)
- [ACDH Vocabularies](#) (Austrian Academy of Sciences)
- [Thesaurus INRAE](#) (Paris, France)
- [AGROVOC Multilingual Thesaurus](#) (United Nations)
- [UNESCO Thesaurus](#)
- [European Space Agency](#) ESA

NDE (Netwerk Digitaal Erfgoed) is working with DANS on the (partial) support of SKOSMOS protocol to get a proper external CV connection to DANS Data Stations.

# Collaboration with GDCC

External controlled vocabulary working group.

Consensus proposal for the CVV support implementation, the current state and requirements matrix:

[https://docs.google.com/document/d/1txdcFuxskRx\\_tLsDQ7KKLFTMR\\_r9IBhorDu3V\\_r445w/edit?ts=607451c0](https://docs.google.com/document/d/1txdcFuxskRx_tLsDQ7KKLFTMR_r9IBhorDu3V_r445w/edit?ts=607451c0)

Pull Request

<https://github.com/IQSS/dataverse/pull/7712>

Demonstration

<http://github.com/GlobalDataverseCommunityConsortium/dataverse/tree/external-cvoc2>



# Known issues with support of external controlled vocabularies

- how CV support could be applied to any field
  - support of any available vocabulary
  - backward compatibility with fields from the old metadata schema
  - clean UI experience (one selection can fill 4 child fields)
  - can we use non-managed vocabularies or free-text values in same field
  - concept drift (the change of meaning of concepts)
  - interoperability across all Dataverse instances
  - how to ensure CVs are coming from authoritative services
- ...

# Issue: how CV support could be applied to any field?

Problem: would support keyword field (with addition of one child field) and any new/existing fields built to have the 4 required child fields. For example, subject, funder ID, grant ID, etc?

Possible solution: changes could be applied to existing text fields without modifying the metadata block, by adding new fields to store URIs. However requires changes on CV interface side.

cvocDemo Metadata ^

<b>cvocDemo</b> ?	<b>Vocabulary</b> ? unesco	<b>Vocabulary URL</b> ? http://skos.um.es/unescothes/CS000	+
	<b>Term</b> ? Social and economic rights	<b>Term URL</b> ? http://skos.um.es/unescothes/C03681	

**Issue: support of any available vocabulary**

**Problem: currently the implementation specific to SKOSMOS protocols which handles many vocabs.**

**Solution: the interface to external API endpoints with vocabularies has been placed outside of Dataverse as external Javascript and could be extended with support of any API, for example, ORCID service.**



# Issue: Backward Compatibility

Problem: our implementation of external controlled vocabularies support requires 4 child fields instead of 3 (default for Dataverse).

Possible solution: create a flyway script to adapt existing fields entries if metadata schema will get extension with new 4th field to keep the concepts URIs. Second option is to keep new field with URIs empty and force depositors to fill it manually.

Keyword ?

Term ?

COVID-19

Vocabulary ?

Wikidata



Vocabulary URL ?

https://www.wikidata.org/wiki/Q8426319

Term ?

Italy

Vocabulary ?

LCSH



Vocabulary URL ?

http://id.loc.gov/authorities/names/n790

Source: [Harvard Dataverse](#)

## Issue: clean User Interface experience

Problem: display retains the 4 fields even though one selection determines all 4. Could hide/disable other fields? With SKOSMOS-served vocabularies, some child fields will be filled automatically.

Possible solution: use more flexible configuration to define 2 child fields (label/URI) instead of 4 where it's possible. Or make 3 fields read-only and managed by Dataverse, not user, if it's unavoidable.

## Issue: non-managed vocabularies or free-text values

Problem: can user mix non-managed controlled vocabularies or free-text values in the same field?

Possible solution: input could allow disabling the selector with some 'manual' mode. If user wants to store some term that doesn't match any entry in CV, it could be allowed to be stored as text. However, it's not sustainable solution - how to sync free-text terms with external CVs?

Issue: how to ensure CV is from an Authoritative service?

Problem: since the service URL is part of config, it could be configured to use other services (a locally managed one, a mirror, etc.)

Possible solution: admin is responsible for the decision to use an authoritative source. However, we don't know how to control this in the distributed network. It could become a serious issue if service is moving from one to another service provider, mirrors should be also considered there.

# Issue: Concept Drift

*Concept drift* is related to the cases where a concept may replace the meaning of other concepts, or other concepts can take over its meaning. Can lead to the problems with data quality, very difficult to trace and address.

Possible scenarios of *concept drift*:

- at the concept identifier level (label drift)
- in the basic properties of the concept (intensional drift)
- to the things the concept refers to (extensional drift)

Source: [Detecting and Reporting Extensional Concept Drift in Statistical Linked Data](#)

Possible solution: create and maintain cache of every concept inside of data repository

## Issue: interoperability across all Dataverse instances

Problem: this implementation requires the same configuration files to import data and metadata from another Dataverse instance. If not configured, shows as 4 child fields by default.

Possible solution: terms from unsupported (undefined) vocabs would just show as their URLs in another instance.

# Required Development for the sustainability

This proposal leverages the work already done in PR [#7712](#). The additional work needed to implement the proposal above includes:

1. Creation of a new vocabulary table (termUri string, term metadata (json text))
  - Column for service type/URL?
  - Column for retrieval date?
2. Add CRUD API for vocabulary table. API would allow addition of a termURI and would then perform a web call to populate the term metadata (versus allowing user input of metadata)
3. Adapt current PR to add termURI to table during upload.
4. Adapt current PR (config file example) to work with a single field versus parent/4 child model
5. Adapt SKOSMOS Javascript to handle display as well as input.
6. Develop plan for migrating existing keyword entries to new model
  - E.g. identify existing CVV entries and identify/create SKOSMOS service to provide them, develop sql script to replace existing values
7. Develop recommendations/documentation/examples to support using CVVs in keyword and custom fields.

# Caching function for CVV

## Linked Data Serverless Resolver as Lambda Function on Harvard AWS cloud

The screenshot shows the GitHub repository page for 'Dans-labs / ld-serverless-resolver'. The repository is owned by '4tikhonov' and has 13 commits. The file list includes folders for 'Id\_proxy', 'scripts', and 'tests', and files for 'Dockerfile', 'LICENSE', 'README.md', 'requirements.txt', 'setup.py', and 'template.yml'. The 'README.md' file is selected and its content is displayed below. The README describes the 'Linked Data Serverless Resolver' as a serverless LD Resolver implemented by DANS R&D as a FastAPI framework, which can be deployed as a Lambda function on Amazon AWS. It also provides links for more information and instructions on how to setup a custom domain name for deployment. On the right side of the repository page, there are sections for 'About', 'Releases', 'Packages', and 'Languages'. The 'Languages' section shows a bar chart with the following data: Python 78.8%, Jupyter Notebook 18.9%, and Dockerfile 2.3%.

File/Folder	Description	Last Commit
Id_proxy	Cleaning up	4 days ago
scripts	Proof of Concept of LD Proxy as Serverless service	4 days ago
tests	Proof of Concept of LD Proxy as Serverless service	4 days ago
Dockerfile	Proof of Concept of LD Proxy as Serverless service	4 days ago
LICENSE	Proof of Concept of LD Proxy as Serverless service	4 days ago
README.md	Motivation and credits	3 days ago
requirements.txt	Requirements updated for local deployment	4 days ago
setup.py	Readme updated with some instructions	4 days ago
template.yml	Proof of Concept of LD Proxy as Serverless service	4 days ago

**Linked Data Serverless Resolver**

Serverless LD Resolver implemented by DANS R&D as FastAPI framework and could be deployed as Lambda function on Amazon AWS. Read more about [AWS Lambda](#), some instructions available how to setup [custom domain name](#) for the deployment.

**Languages**

- Python 78.8%
- Jupyter Notebook 18.9%
- Dockerfile 2.3%

### Features:

- Shared service for all Dataverse instances
- Memento protocol support must have
- Integrated with LD Proxy service
- Archived concepts for every dataset version

<https://github.com/Dans-labs/ld-serverless-resolver>



# Caching concept URIs

## WikiData

```
{
  term: "Q82069695",
  - json: {
    - entities: {
      - Q82069695: {
        pageid: 81427064,
        ns: 0,
        title: "Q82069695",
        lastrevid: 1432714450,
        modified: "2021-05-31T23:58:40Z",
        type: "item",
        id: "Q82069695",
        - labels: {
          - de: {
            language: "de",
            value: "SARS-CoV-2"
          },
          - en: {
            language: "en",
            value: "SARS-CoV-2"
          },
          - zh: {
            language: "zh",
            value: "严重急性呼吸系统综合征冠状病毒2"
          },
          - th: {
            language: "th",
            value: "ไวรัสโคโรนาสายพันธุ์ใหม่ (SARS-CoV-2)"
          },
          - zh-hans: {
            language: "zh-hans",
            value: "严重急性呼吸系统综合征冠状病毒2"
          },
          - es: {
            language: "es",
            value: "SARS-CoV-2"
          },
          - zh-cn: {
            language: "zh-cn",
            value: "严重急性呼吸系统综合征冠状病毒2"
          },
          - wuu: {
            language: "wuu",
            value: "SARS-CoV-2"
          },
        }
      }
    }
  }
}
```

## MeSH

```
{
  term: "D045473",
  - json: {
    @id: "http://id.nlm.nih.gov/mesh/D045473",
    @type: "http://id.nlm.nih.gov/mesh/vocab#TopicalDescriptor",
    http://id.nlm.nih.gov/mesh/vocab#active: true,
    - allowableQualifier: [
      "http://id.nlm.nih.gov/mesh/Q000276",
      "http://id.nlm.nih.gov/mesh/Q000378",
      "http://id.nlm.nih.gov/mesh/Q000502",
      "http://id.nlm.nih.gov/mesh/Q000302",
      "http://id.nlm.nih.gov/mesh/Q000472",
      "http://id.nlm.nih.gov/mesh/Q000145",
      "http://id.nlm.nih.gov/mesh/Q000187",
      "http://id.nlm.nih.gov/mesh/Q000737",
      "http://id.nlm.nih.gov/mesh/Q000235",
      "http://id.nlm.nih.gov/mesh/Q000648",
      "http://id.nlm.nih.gov/mesh/Q000528",
      "http://id.nlm.nih.gov/mesh/Q000254",
      "http://id.nlm.nih.gov/mesh/Q000201"
    ],
    - annotation: {
      @language: "en",
      @value: "infection = SEVERE ACUTE RESPIRATORY SYNDROME"
    },
    broaderDescriptor: "http://id.nlm.nih.gov/mesh/D000073640",
    dateCreated: "2003-04-17",
    dateEstablished: "2003-04-17",
    dateRevised: "2020-08-13",
    - historyNote: {
      @language: "en",
      @value: "2003"
    },
    identifier: "D045473",
    nlmClassificationNumber: "QW 168.5.C8",
    preferredConcept: "http://id.nlm.nih.gov/mesh/M0448382",
    preferredTerm: "http://id.nlm.nih.gov/mesh/T538594",
    - publicMeSHNote: {
      @language: "en",
      @value: "2003"
    },
    treeNumber: "http://id.nlm.nih.gov/mesh/B04.820.578.500.540.150.113.937",
    - label: {
      @language: "en",
      @value: "SARS Virus"
    }
  }
}
```

Archived concepts incorporated in the dataset metadata export is the link to Linked Open Data!

# Linking data (files) to external CVs, not only metadata

Forum Research Political Poll - Federal Issues (Canada) 2015

FOCN\_SPSS\_20150525\_FORMATTED.tab

Forum Research Inc., 2016, "Forum Research Political Poll - Federal Issues (Canada) 2015", <https://hdl.handle.net/10864/12144>, Scholars Portal Dataverse, V1

< Hide Groups

Add Group + Search

All Variables

ID	Name	Weight	View
<input type="checkbox"/>	v134374 Q1A		
<input type="checkbox"/>	v134345 Q1B		
<input type="checkbox"/>	v134615 Q2		
<input type="checkbox"/>	v134478 Q3		
<input type="checkbox"/>	v134600 Q4		
<input type="checkbox"/>	v134634 Q5		
<input type="checkbox"/>	v134338 Q6		
<input type="checkbox"/>	v134522 Q7		
<input type="checkbox"/>	v134310 Q8		
<input type="checkbox"/>	v134373 Q9		
<input type="checkbox"/>	v134658 Q10		
<input type="checkbox"/>	v134377 Q11A		
<input type="checkbox"/>	v134381 Q11B		
<input type="checkbox"/>	v134678 Q11C		
<input type="checkbox"/>	v134340 Q12A		

### Variable Information

ID: v134634 Name: Q5

Label: 5. Which party do you expect to win the next federal election?

Literal Question: \_\_\_\_\_

Interviewer Instructions: \_\_\_\_\_

Post Question: \_\_\_\_\_

Universe: \_\_\_\_\_

Notes: \_\_\_\_\_

Group: \_\_\_\_\_

Weight Variable:  Is Weight

Update Cancel

Source: Scholars Portal' [Data Curation Tool](#) (Canada)

Thank you for your attention!

Slava Tykhonov (DANS-KNAW)

[vyacheslav.tykhonov@dans.knaw.nl](mailto:vyacheslav.tykhonov@dans.knaw.nl)

Join our community



<https://www.sshopencloud.eu>



@SSHOpenCloud



[info@sshopencloud.eu](mailto:info@sshopencloud.eu)



/in/sshopencloud

