

# Metadata Schemas in Dataverse

Dataverse Community Meeting 2021  
June 14 to 17, 2021

Breakout Session on Flexible Metadata and Controlled Vocabularies

Philipp Konzett  
UiT The Arctic University of Norway  
ORCID: [0000-0002-6754-7911](https://orcid.org/0000-0002-6754-7911)  
Twitter: @PhilippKonzett

John Graybeal  
Stanford University  
<https://profiles.stanford.edu/john-graybeal>



# Outline

1. Introduction and current work on improvements (Philipp)
2. Some ideas about how to further enhance metadata schema support in Dataverse (Philipp & John)
3. Q&A and discussion

1.

Introduction and  
current work on improvements

# Short recap from Flexible Metadata Breakout Session at Dataverse Community Meeting 2020

- ❑ Summary of desirable properties of metadata support in Dataverse
- ❑ Based on GitHub issues and discussions in Google Group
- ❑ Recap of a few selected desirables

# Desirables for Dataverse Metadata Support

## COVERAGE

- ❑ **Subjects:** metadata cover **as many subjects/domains as relevant** for the Dataverse community (requested domain coverage includes biology, chemistry, code-based data, linguistics, physics)

## COMPLIANCE AND CONTROL

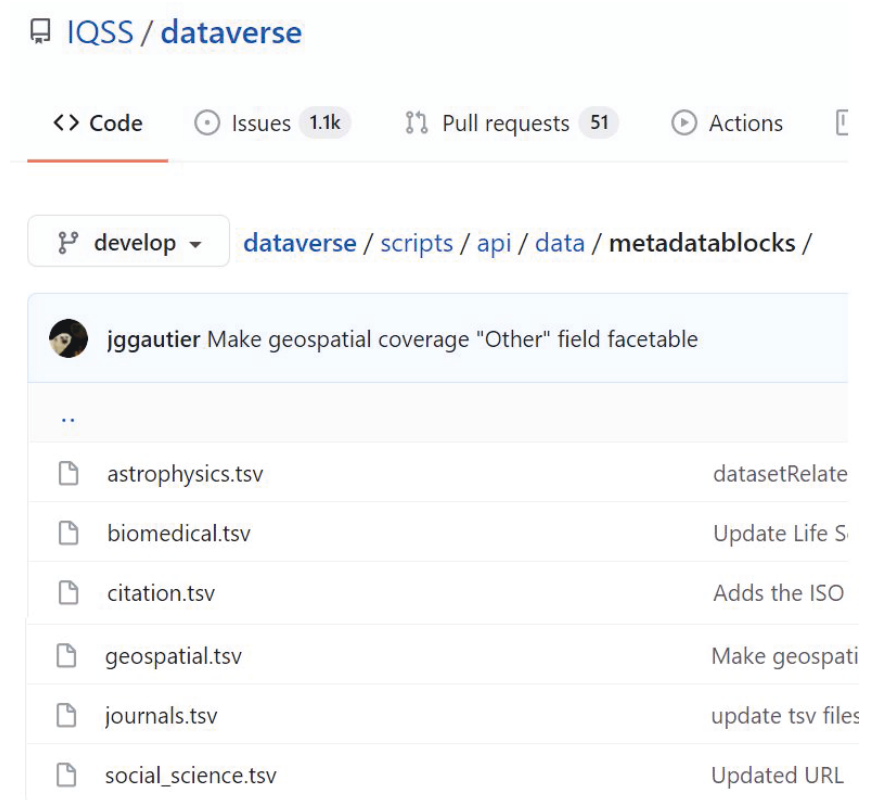
- ❑ **Standards:** metadata are **compliant with standards**
- ❑ **Controlled vocabularies:** the values of fields are determined by **external, controlled, and standardized vocabularies**, which preferably have internationalization support
- ❑ **Export:** metadata are provided for **harvesting and export in compliant formats**

## INTEGRATION

- ❑ **Integration:** metadata schemas allow seamless **integration with external tools and services**, e.g. controlled vocabularies, curation tools, ELNs, preservation tools, etc.  
**+ metadata management tools like CEDAR ...**

# Current metadata schema support in Dataverse

- Metadata schemas are defined in **tsv files**.
- The general metadata schema (citation metadata) and a few domain-specific schemas are shipped with the **main distribution** of Dataverse.
- Customized schemas can be uploaded by repository admin.



The screenshot shows the GitHub interface for the 'dataverse' repository. At the top, it displays the repository name 'IQSS / dataverse' and navigation links for 'Code', 'Issues' (1.1k), 'Pull requests' (51), and 'Actions'. Below this, the current branch is 'develop' and the file path is 'dataverse / scripts / api / data / metadatablocks /'. A pull request by user 'jggautier' is shown with the title 'Make geospatial coverage "Other" field facetable'. Below the pull request, a table lists several TSV files and their associated actions:

..	
astrophysics.tsv	datasetRelate
biomedical.tsv	Update Life S
citation.tsv	Adds the ISO
geospatial.tsv	Make geospati
journals.tsv	update tsv files
social_science.tsv	Updated URL

# Two examples of customized metadata schemas being developed in the community

- ❑ Use Case 1: CESSDA Metadata Model (CMM)
- ❑ Use Case 2: Component Metadata Infrastructure (CMDI)

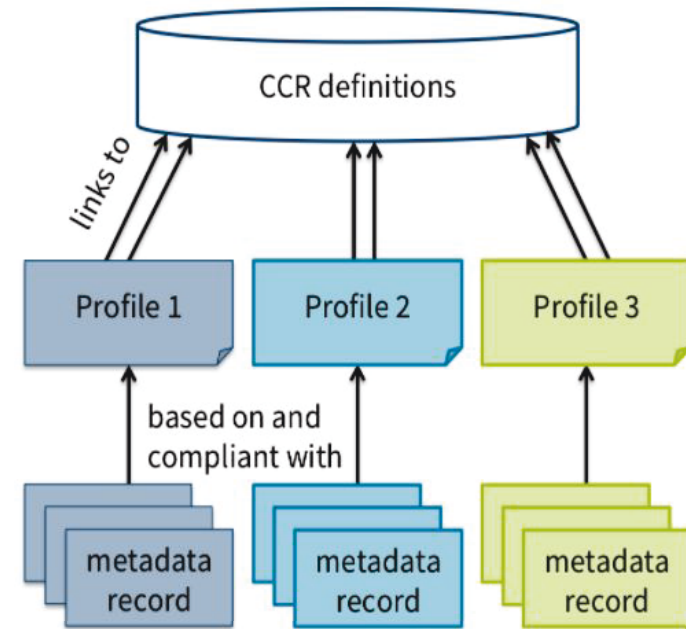
# Use Case 1: CESSDA Metadata Model (CMM)

- ❑ Developed by the Metadata Office of CESSDA (Consortium of European Social Science Data Archives), which is a European Research Infrastructure Consortium (ERIC). Main objective is to **standardize metadata for CESSDA Service Providers**.
- ❑ Built from the viewpoint of **quantitative (social science) data** and based on the DDI Lifecycle 3.2 metadata standard.
- ❑ Contains **metadata elements**, their **definitions** and other **requirements**, such as repeatability and use of certain Controlled Vocabularies.
- ❑ Consists of **11 top elements** with information about (1) Study, (2) Person(s), (3) Organizations, (4) Dataset, (5) Instrument, (6) Questions and Answers, (7) Concepts, (8) Further Documents, (9) Publications, (10) Group of Studies, (11) Document Description ('metadata about metadata').
- ❑ **DANS** is working in the **SSHOC project** on **implementing** the CESSDA Metadata Model in the **Dataverse** software. Issues found so far:
  - ❑ Dataverse does not support language attributes;
  - ❑ Social Sciences block should be updated to enable the use of Controlled Vocabularies;
  - ❑ Saving variable metadata is limited.



# Use Case 2: Component MetaData Infrastructure (CMDI)

- ❑ CMDI is developed by **CLARIN** (Common Language Resources and Technology Infrastructure; = an ERIC), and is used as a standard for metadata provision from CLARIN centres.
- ❑ CMDI is a **framework** to describe and reuse metadata blueprints.
- ❑ **Profiles** consist of **components** which may be based on multiple metadata standards (e.g. Dublin Core, DDI Codebook, etc.).
- ❑ Components are registered in the CMDI **Component Registry (CCR)**.
- ❑ CLARIN is working on a recommendation for **core metadata**.
- ❑ DANS (Easy and DataverseNL) and UiT (TROLLing) are working in the **SSHOC project** on implementing **CMDI support in Dataverse**, including recommendations for CLARIN core metadata.



(Image: <https://www.clarin.eu/content/component-metadata>)

2.

Some ideas about how to further enhance metadata schema support in Dataverse

# First: Why enhance metadata support in Dataverse?

1. Make metadata used in Dataverse **more interoperable**
2. Make metadata support **more sustainable**

FLEXIBLE

INTEROPERABLE

biology\_md\_repo\_A.tsv

⌘?

darwin\_core\_md\_repo\_A.tsv

↕?

biology\_md\_repo\_B.tsv

darwin\_core\_md\_repo\_B.tsv

- No standard
- High flexibility
- Low interoperability

- Based on standard
- Less flexibility
- ?Different implementation
- ?Medium interoperability

INTEGRATION

“It is pivotal to guide the data producers to **capture** essential **metadata** elements **from the onset**, rather than upon submission to a repository, find out that some metadata field is missing, or that a specific controlled vocabulary was required” (Wittenburg et al. 2019)

? SUSTAINABLE

# How to achieve higher interoperability and sustainability?

On a practical level, we could start working in **two stages**:

- **Stage 1 -- based on tsv files:** Include requested metadata schemas as tsv files shipped with **main distribution** of Dataverse software
- **Stage 2 -- through integration with external tools, e.g. CEDAR:** Support multiple metadata standards as well as flexible metadata through integration with Center for Expanded Data Annotation and Retrieval (CEDAR)

# Stage 1 – Enhancing tsv file based support

- Metadata schemas are defined in **tsv files**.
- The general metadata schema (citation metadata) and a few domain-specific schemas are shipped with the **main distribution** of Dataverse.
- Customized schemas can be uploaded by repository admin.
- **Agree on and include more (domain-specific) schemas, e.g. Linguistic metadata, Darwin Core, in the main distribution of Dataverse.**

IQSS / dataverse

<> Code

! Issues 1.1k

🔗 Pull requests 41

▶ Actions

📁 Projects

🔗 develop ▾

dataverse / scripts / api / data / metadatablocks /

..

📄 astrophysics.tsv

📄 biomedical.tsv

📄 citation.tsv

📄 linguistics.tsv

📄 biology.tsv

...

# Why should we go beyond stage 1?

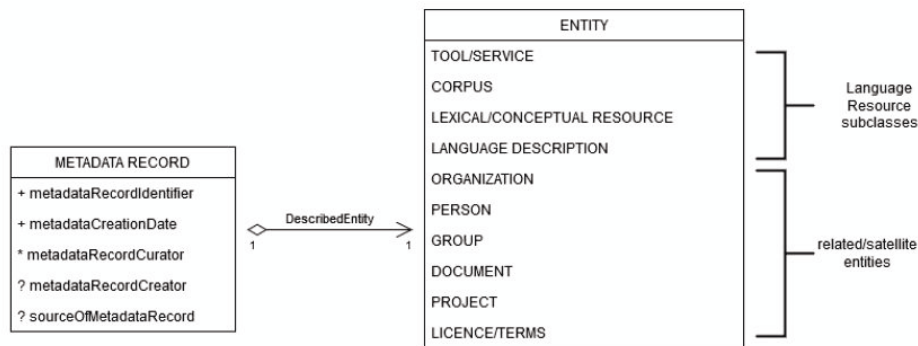
Tsv approach has its **limitations**, e.g. when dealing with

- **hierarchical** and more **complex** metadata standards;
- **integration** with tools used in the active phase of (meta)data management.

Also: Translating large and complex metadata standards into csv files is a laborious process.

Example: European Language Grid (ELG) Metadata Schema ...

The following figure shows the main notions upon which the ELG schema builds.

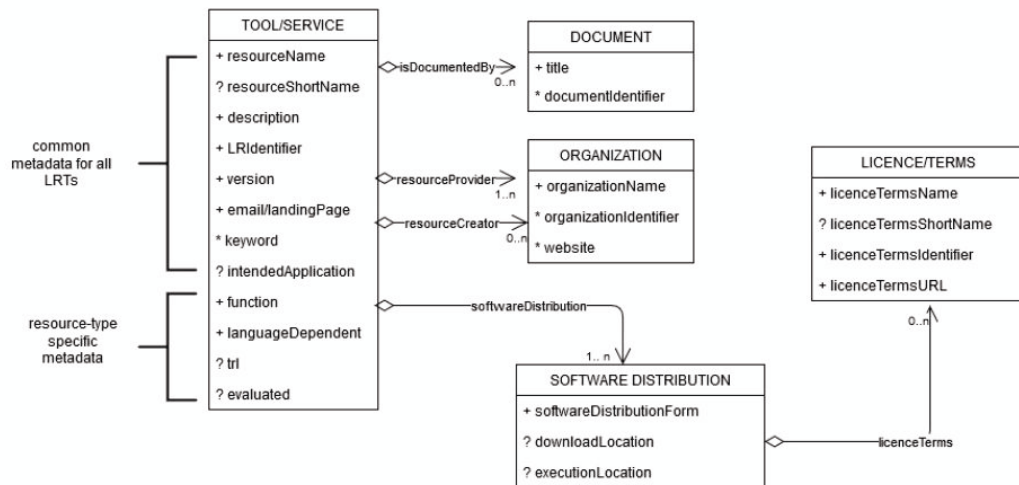


# Conceptual outline of the European Language Grid (ELG) Metadata Schema

A CMDI compliant metadata schema

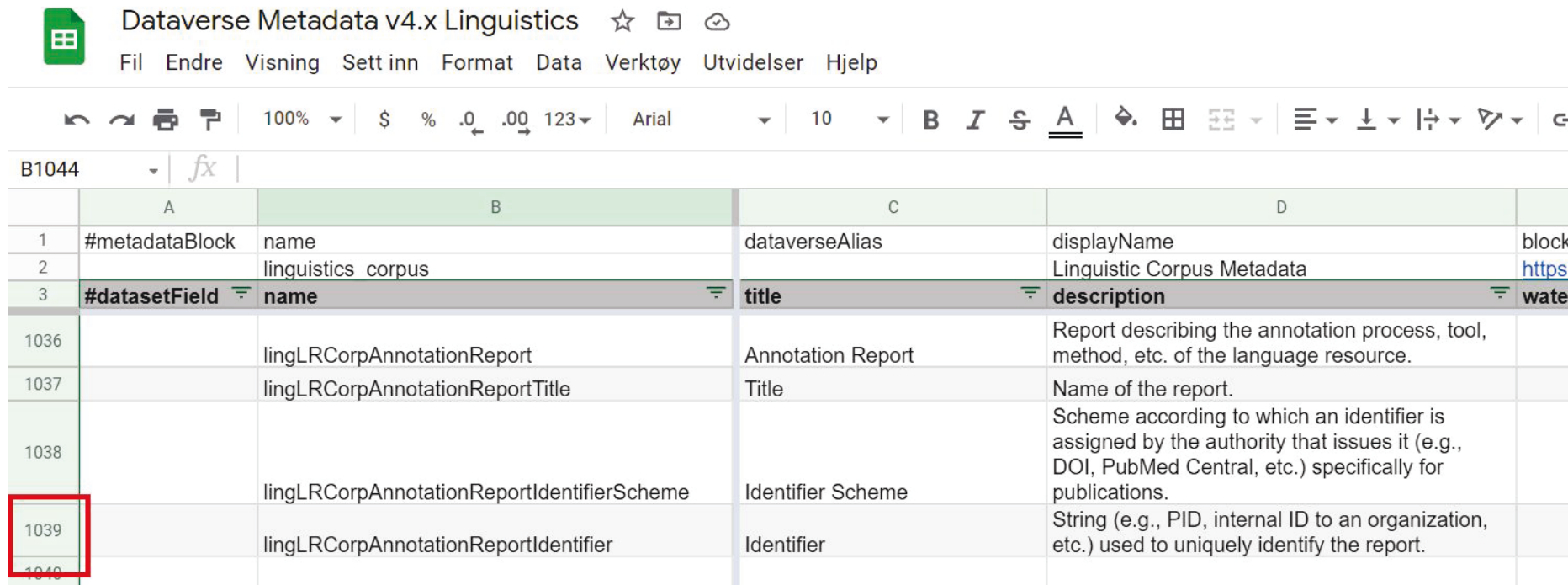
(Image:

[https://european-language-grid.readthedocs.io/en/release1.1.1/all/A1\\_Metadata/Metadata.html](https://european-language-grid.readthedocs.io/en/release1.1.1/all/A1_Metadata/Metadata.html))





... one component may look like this in a tsv file (Google spreadsheet):



Dataverse Metadata v4.x Linguistics ☆ 📁 ☁

Fil Endre Visning Sett inn Format Data Verktøy Utvidelser Hjelp

100% \$ % .0 .00 123 Arial 10 B I S A 🔍 📄 📑 ⌵ ⌴ ↗

B1044 | fx

	A	B	C	D	
1	#metadataBlock	name	dataverseAlias	displayName	block
2		linguistics corpus		Linguistic Corpus Metadata	<a href="#">https</a>
3	#datasetField	name	title	description	wate
1036		lingLRCorpAnnotationReport	Annotation Report	Report describing the annotation process, tool, method, etc. of the language resource.	
1037		lingLRCorpAnnotationReportTitle	Title	Name of the report.	
1038		lingLRCorpAnnotationReportIdentifierScheme	Identifier Scheme	Scheme according to which an identifier is assigned by the authority that issues it (e.g., DOI, PubMed Central, etc.) specifically for publications.	
1039		lingLRCorpAnnotationReportIdentifier	Identifier	String (e.g., PID, internal ID to an organization, etc.) used to uniquely identify the report.	
1040					

The number of rows may indicate the size and complexity...

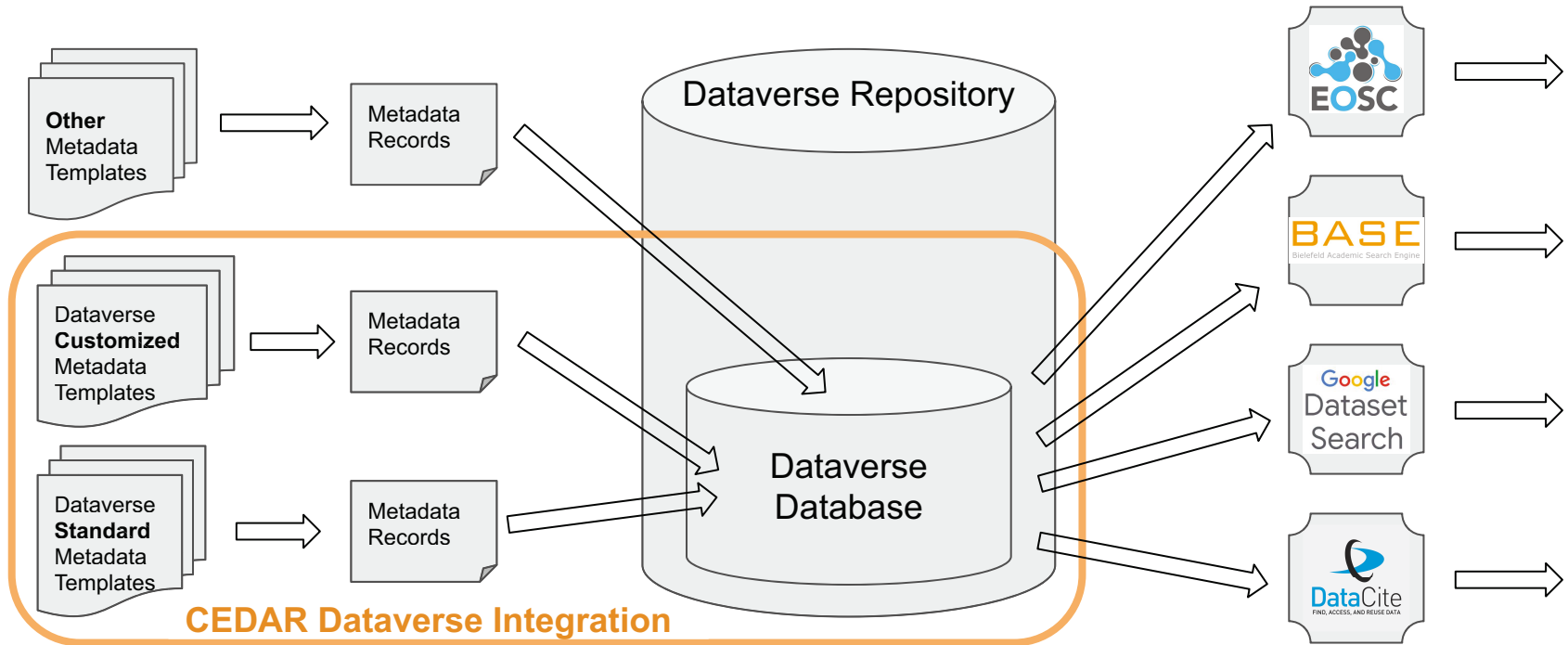
Are there other ways to support metadata standards in Dataverse?

## Stage 2 – CEDAR integration: Basic idea

- Two types of metadata schemas in Dataverse:
  - a) **standard** schemas: shipped with main distribution
  - b) **customized** schemas: created by Dataverse installation
- Define and maintain Dataverse standard and customized metadata schemas **as CEDAR templates**.
- Get metadata into Dataverse through
  - a) registration in integrated CEDAR template, or
  - b) upload of CEDAR metadata file
- In addition: Support upload of researcher-customized CEDAR metadata files for more granular metadata not covered by standard or customized schemas in Dataverse.



## Export / Harvesting






Planning    ...

Project management  ...

Discovery → Collection → Processing → Analyzing → Sharing


   ...

Documentation    ... 

Standards, Databases, Policies  ...

Reuse

biology\_md\_repo\_A.tsv

⌘ ?

darwin\_core\_md\_repo\_A.tsv

↕ ?

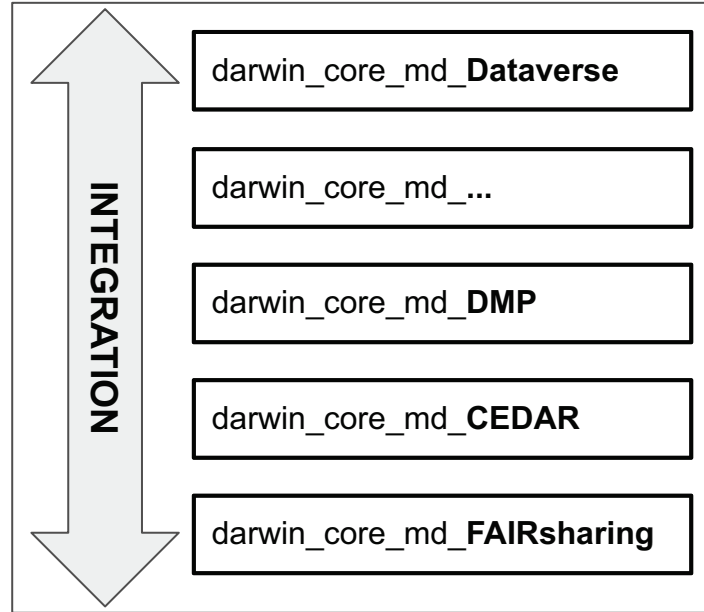
biology\_md\_repo\_B.tsv

darwin\_core\_md\_repo\_B.tsv

darwin\_core\_md\_dataverse.tsv

+ CVV

“It is pivotal to guide the data producers to capture essential metadata elements from the onset, rather than upon submission to a repository, find out that some metadata field is missing, or that a specific controlled vocabulary was required” (Wittenburg et al. 2019)



# CEDAR: End to End Metadata



CEDAR

CENTER FOR EXPANDED DATA  
ANNOTATION AND RETRIEVAL

metadacenter.org

<https://metadacenter.org/references>



## MiAIRR

### BioProject

Study ID\*

PRJNA429427

Study Title\*

DN B cells in Multiple Sclerosis (MS) patients and healthy controls (HC)

Study Type\*

AIRR-seq

### Study Criteria

MS patients were recruited at the Revalidation & MS Center (Overpelt, Belgium) and were diagnosed according to the McDonald criteria. HC were recruited at the Biomedical Research Institute (Diepenbeek, Belgium).

Funding Agency\*

This work was supported by the National Multiple Sclerosis Society (PP-1509-06370), the Belgian Charcot Foundation and the National Institutes of Health (NIH) (R01AI104739).



# Some Important considerations

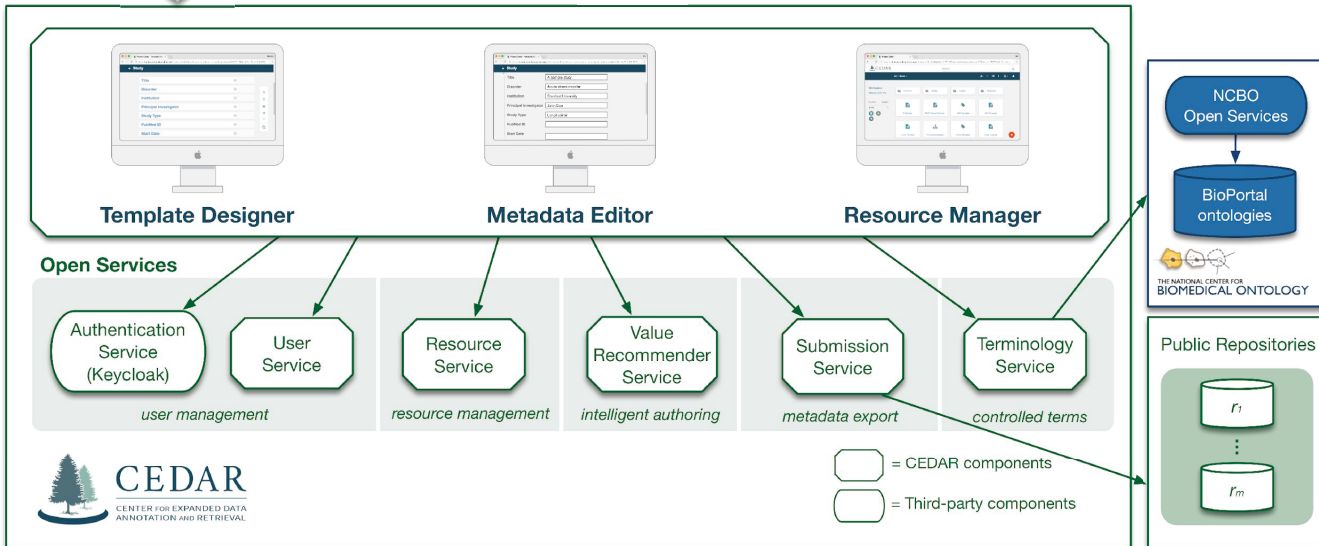
- Available tooling to work with the metadata in a community way
- Semantics (linked data identifiers for concepts)
- Mature, powerful, & flexible metadata specification (hierarchy, semantics, ...)
- Reusable, modular, trusted metadata specifications: 'Dataverse-approved'



**Template Specifiers**



**Research Scientists**



# CEDAR Workspace elements



Template fields



Template elements  
(groups of fields and/or elements)



Metadata templates

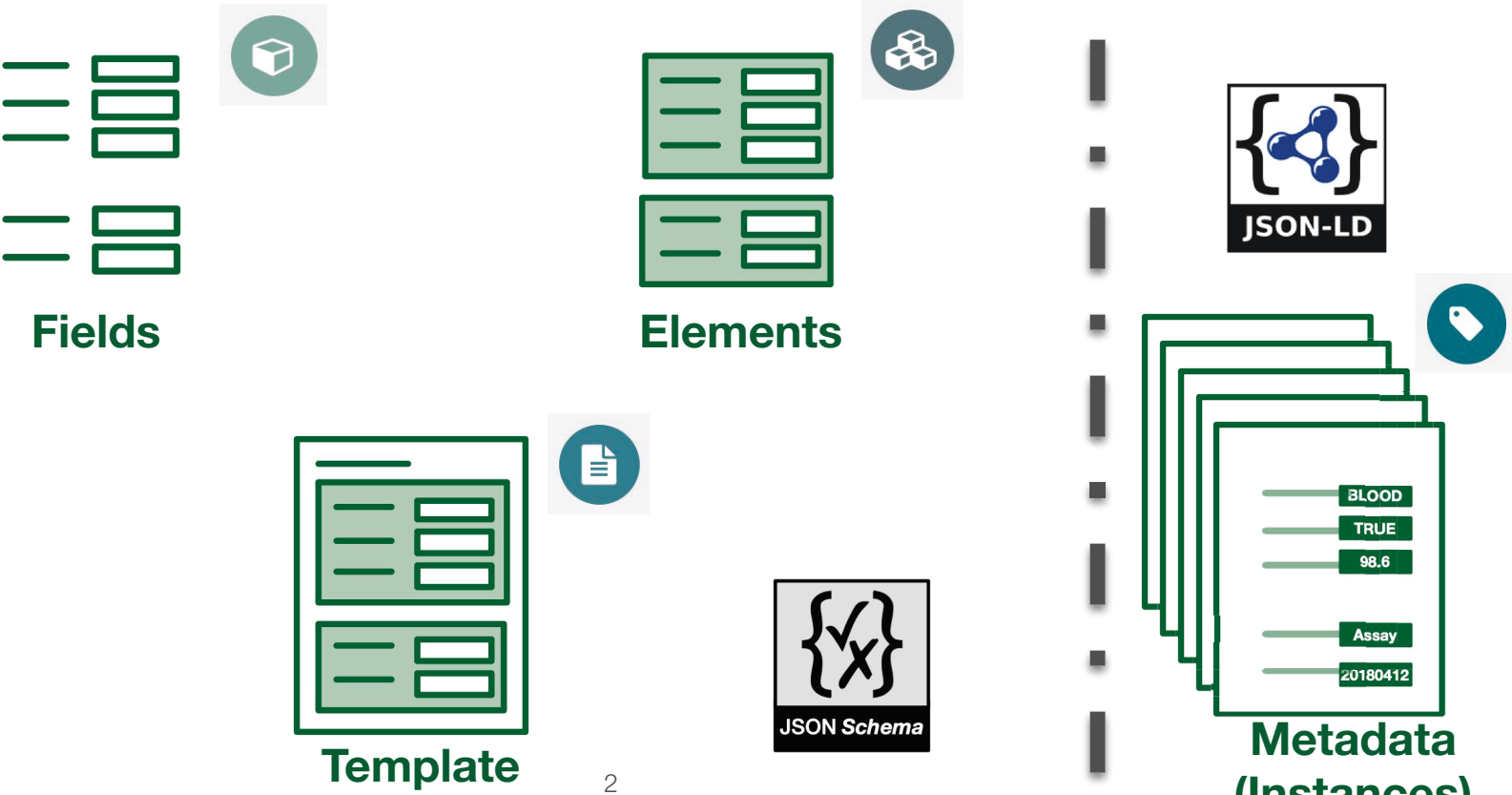


Instantiated metadata



Folders

# CEDAR's Template Model



# Model Specifications



PURPOSE RESEARCH TOOLS | TRAINING COMMUNITY

Better data for better science

## CEDAR Template Model V1.5.0

Martin J O'Connor, Marcos Martínez-Romero, and the

CEDAR Team

October 15th, 2018

[Home](#)

## CEDAR Template Model V1.5.0

One of the main goals of the CEDAR project is to build an infrastructure for the creation and storage of templates. These templates provide detailed definitions of the metadata that describes a particular data resource. (the data resource) may be described with metadata in a comma-separated values (CSV) file. The template specification applies to other CSV files following the same specification.

This report describes the format of the templates, elements, and instances that make up the CEDAR metadata to be used by CEDAR, and exchanged with other metadata systems.

We also have written [a paper describing the model](#), which provides a more formal written description.

**Release Date:** October 15, 2018

**Attachment:**  [CEDAR Template Model v1.5.0.pdf](#)

**Author List:** Martin O'Connor, Marcos Martinez-Romero, and the CEDAR team

**Artifact Type:** [Report](#)

## **An Open Repository Model for Acquiring Knowledge about Scientific Experiments**

Martin J. O'Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Willrett, John Graybeal, and Mark A. Musen

Stanford Center for Biomedical Informatics Research  
Stanford, CA 94305, U.S.A.  
[martin.oconnor@stanford.edu](mailto:martin.oconnor@stanford.edu)

**Abstract.** The availability of high-quality metadata is key to facilitating discovery in the large variety of scientific datasets that are increasingly becoming publicly available. However, despite the recent focus on metadata, the diversity of metadata representation formats and the poor support for semantic markup typically result in metadata that are of poor quality. There is a pressing need for a metadata representation format that provides strong interoperability capabilities together with robust semantic underpinnings. In this paper, we describe such a format, together with open-source Web-based tools that support the acquisition, search, and management of metadata. We outline an initial evaluation using metadata from a variety of biomedical repositories.

New +

### Workspace

Shared with Me

FILTER

TYPE














VERSION

Latest

All / MiAIRR



Title	Crea...	Modi...	
 MiAIRR	3:35 PM	3:35 PM	 
 BioSample	3:34 PM	3:34 PM	
 BioProject	3:34 PM	3:34 PM	
 SRA	3:34 PM	3:34 PM	
 DN B cells in Multiple Sclerosis	3:28 PM	3:30 PM	

Displaying 1 - 5 of 5

New +

All / MiAIRR



## Workspace

Shared with Me

FILTER













TYPE



VERSION



Latest

Title	Crea...	Modi...	
 MiAIRR	3:35 PM	3:35 PM	 
 BioSample	3:34 PM	3:34 PM	
 BioProject	3:34 PM	3:34 PM	
 SRA	3:34 PM	3:34 PM	
  n Multiple Sclerosis	3:28 PM	3:30 PM	

Displaying 1 - 5 of 5



## MiAIRR

### BioProject

Study ID\*

PRJNA429427

Study Title\*

DN B cells in Multiple Sclerosis (MS) patients and healthy controls (HC)

Study Type\*

AIRR-seq

Study Criteria

MS patients were recruited at the Revalidation & MS Center (Overpelt, Belgium) and were diagnosed according to the McDonald criteria. HC were recruited at the Biomedical Research Institute (Diepenbeek, Belgium).

Funding Agency\*

This work was supported by the National Multiple Sclerosis Society (PP-1509-06370), the Belgian Charcot Foundation and the National Institutes of Health (NIH) (R01AI104739).














New +

- Folder
- Field
- Element
- Template

All / MiAIRR



Title	Crea...	Modi...	
 MiAIRR	3:35 PM	3:35 PM	 
 BioSample	3:34 PM	3:34 PM	
 BioProject	3:34 PM	3:34 PM	
 SRA	3:34 PM	3:34 PM	
 DN B cells in Multiple Sclerosis	3:28 PM	3:30 PM	



FILTER

TYPE



VERSION

Displaying 1 - 5 of 5

Name

MiARR

Identifier

Description



CANCEL

SAVE TEMPLATE



{ } JSON Schema





Name

MiARR

Identifier

Description



Enter Field Name

Submission Release Date



Enter Preferred Label

Enter Field Help Text

The date (in yyyy-mm-dd format) upon which all the submissions in this metadata fo

MULTIPLE

REQUIRED



CLEAR

CANCEL

SAVE TEMPLATE

Name

MiARR

Identifier

Description



✚ 📅 ✎ ✕

Enter Field Name  
**Submission Release Date** 📎

Enter Preferred Label

Enter Field Help Text  
The date (in yyyy-mm-dd format) upon which all the submissions in this metadata for

MULTIPLE REQUIRED

- A
- 📅
- ✉
- #
- ⋮
- 🔍





Submissions Release Date



BioProject for AIRR NCBI



Study ID



Study Title



Study Type



Study Criteria



Funding Agency



A



#







Name


MiAIRR

Identifier

Description





   

Enter Field Name  
**Submission Release Date** 

Enter Preferred Label

Enter Field Help Text  
The date (in yyyy-mm-dd format) upon which all the submissions in this metadata fo

MULTIPLE REQUIRED

- 
- 
- 
- 
- 
- 

CLEAR

38  
CANCEL

SAVE TEMPLATE




Name

MiARR

Identifier

Description



Enter Field Name  
**Submission Release Date** 

Enter Preferred Label

Enter Field Help Text  
The date (in yyyy-mm-dd format) upon which all the submissions in this metadata for

MULTIPLE REQUIRED



- A**
- 
- 
- #**
- ⋮
- 



Study Group Description



+ A \* ✎



Enter Field Name

**Disease**



Enter Field Help Text

Enter Default Value

**VALUES**

MULTIPLE

REQUIRED

SUGGESTIONS

HIDDEN

**Name**

**Type**

**Source**

**Identifier**

**No. Values**

disease

Branch

DOID

DOID\_4

-



SEARCH



+ A \* ✎



Length of Disease

40







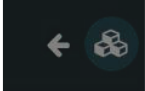
[Start Over](#)

Search in BioPortal  
 disease  

500 results for the query 'disease'. Click on a term below to select it

TERM	DEFINITION	TYPE	SOURCE	ID
	any of...			
disease	-	Class	ICPC2P	A99001
disease	A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more...	Class	DOID	DOID_4
disease	A disease is a disposition that describes states of disease associated with a particular sample and/or	Class	EFO	EFO_0000408





symptoms. It may affect the whole body or any of...				
disease	-	Class	ICPC2P	A99001
disease	A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more...	Class	DOID	DOID_4 <a href="#">Hide Details</a>
disease	A disease is a disposition that describes states of disease associated with a particular sample and/or organism.	Class	EFO	EFO_0000408

**DOID classes:**

- [-] Disease
  - [+] Syndrome
  - [+] Disease Of Cellular Proliferation
  - [+] Genetic Disease
  - [+] Disease Of Anatomical Entity
  - [+] Disease By Infectious Agent
  - [+] Disease Of Metabolism
  - [+] Disease Of Mental Health
  - [+] Physical Disorder

TERM DETAILS

ONTOLOGY DETAILS

<b>Name</b>	disease
<b>Id</b>	<a href="http://purl.obolibrary.org/obo/DOID_4">http://purl.obolibrary.org/obo/DOID_4</a>
<b>Definition</b>	A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.





## DOID classes:

- [-] Disease
  - [+] Syndrome
  - [+] Disease Of Cellular Proliferation
  - [+] Genetic Disease
  - [+] Disease Of Anatomical Entity
  - [+] Disease By Infectious Agent
  - [+] Disease Of Metabolism
  - [+] Disease Of Mental Health
  - [+] Physical Disorder

### TERM DETAILS

### ONTOLOGY DETAILS

<b>Name</b>	disease
<b>Id</b>	<a href="http://purl.obolibrary.org/obo/DOID_4">http://purl.obolibrary.org/obo/DOID_4</a>
<b>Definition</b>	A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.

TERM

**BRANCH**

ONTOLOGY

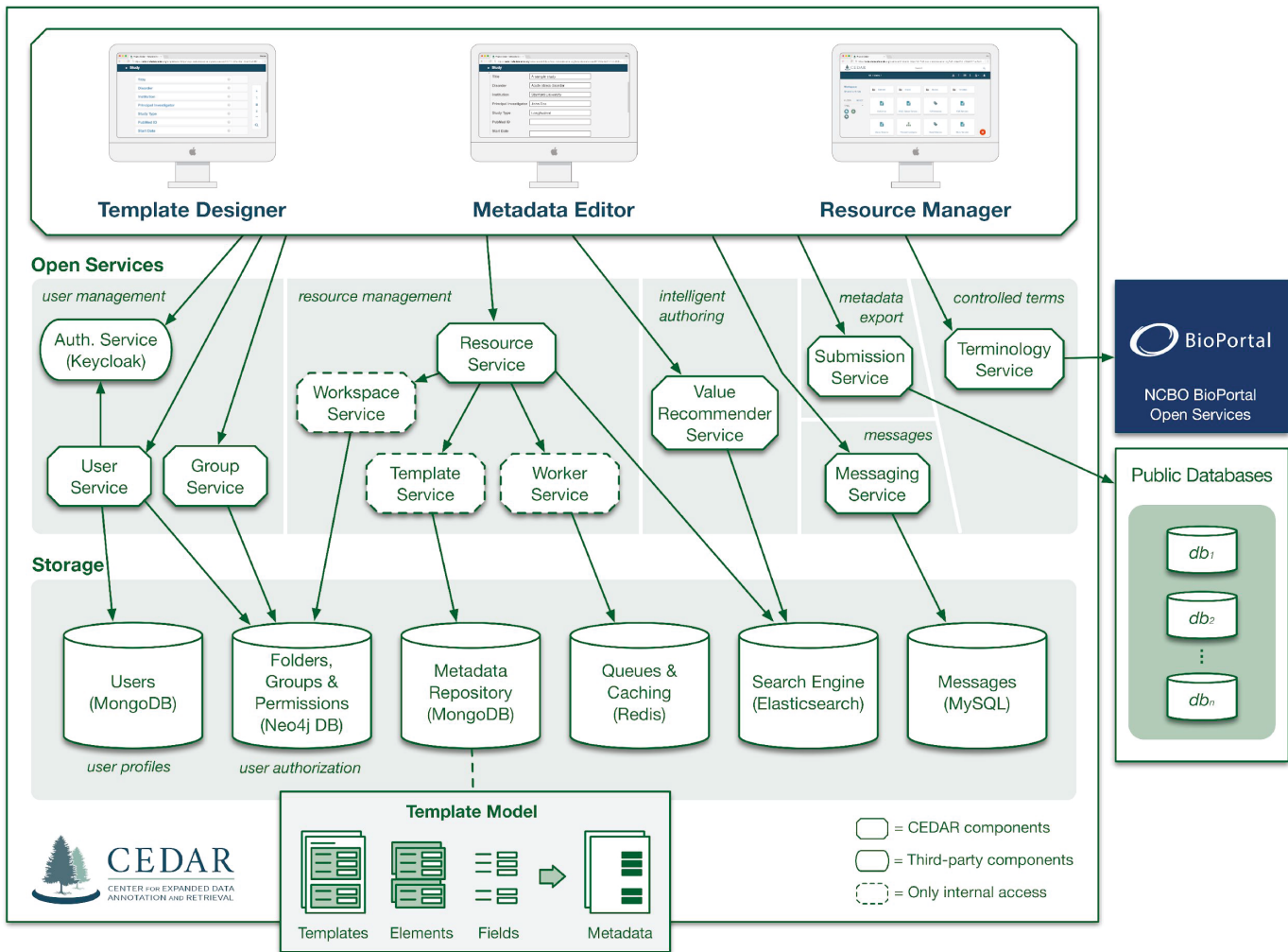
**Term Id** [http://purl.obolibrary.org/obo/DOID\\_4](http://purl.obolibrary.org/obo/DOID_4)

**Term Name** disease

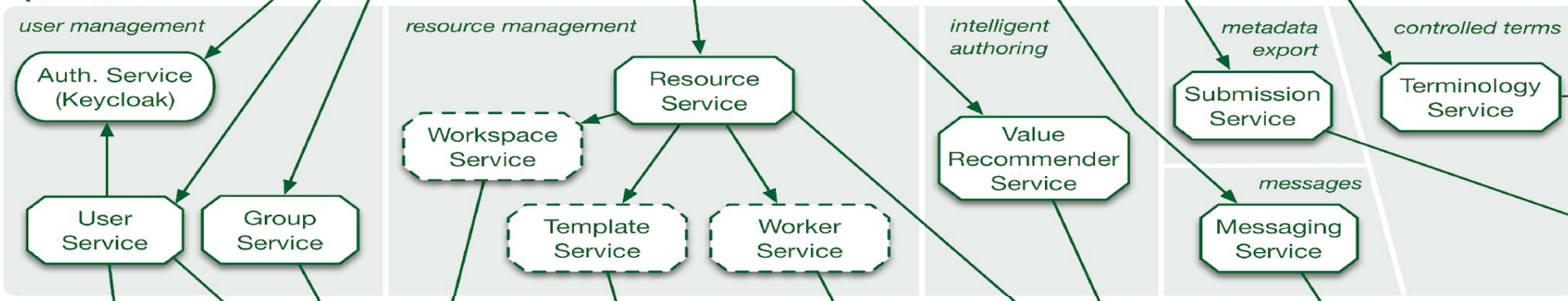
Click to add all the descendants of the selected term

ADD  
43





# Open Services



## CEDAR Resource Server API

Better Metadata - Better Science

**Users** [Show/Hide](#) [List Operations](#) [Expand Operations](#)

**Folders** [Show/Hide](#) [List Operations](#) [Expand Operations](#)

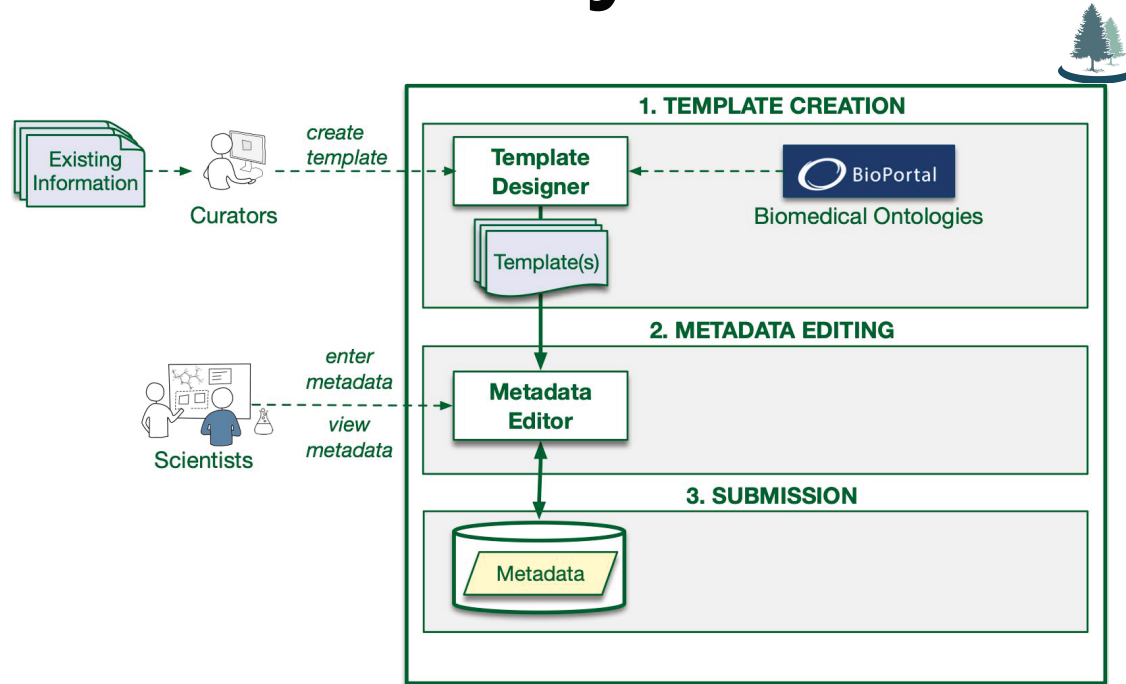
<b>DELETE</b>	/folders/{folder_id}	Delete a folder
<b>GET</b>	/folders/{folder_id}	Get a folder
<b>PUT</b>	/folders/{folder_id}	Update a folder
<b>POST</b>	/folders	Create a folder
<b>GET</b>	/folders/{folder_id}/permissions	Get permissions of a folder
<b>PUT</b>	/folders/{folder_id}/permissions	Update permissions of a folder
<b>GET</b>	/folders/{folder_id}/details	Get the details of a folder
<b>GET</b>	/folders/{folder_id}/contents	Get the contents of a folder
<b>GET</b>	/search	Search resources

**Implementation Notes**  
Search resources using different criteria

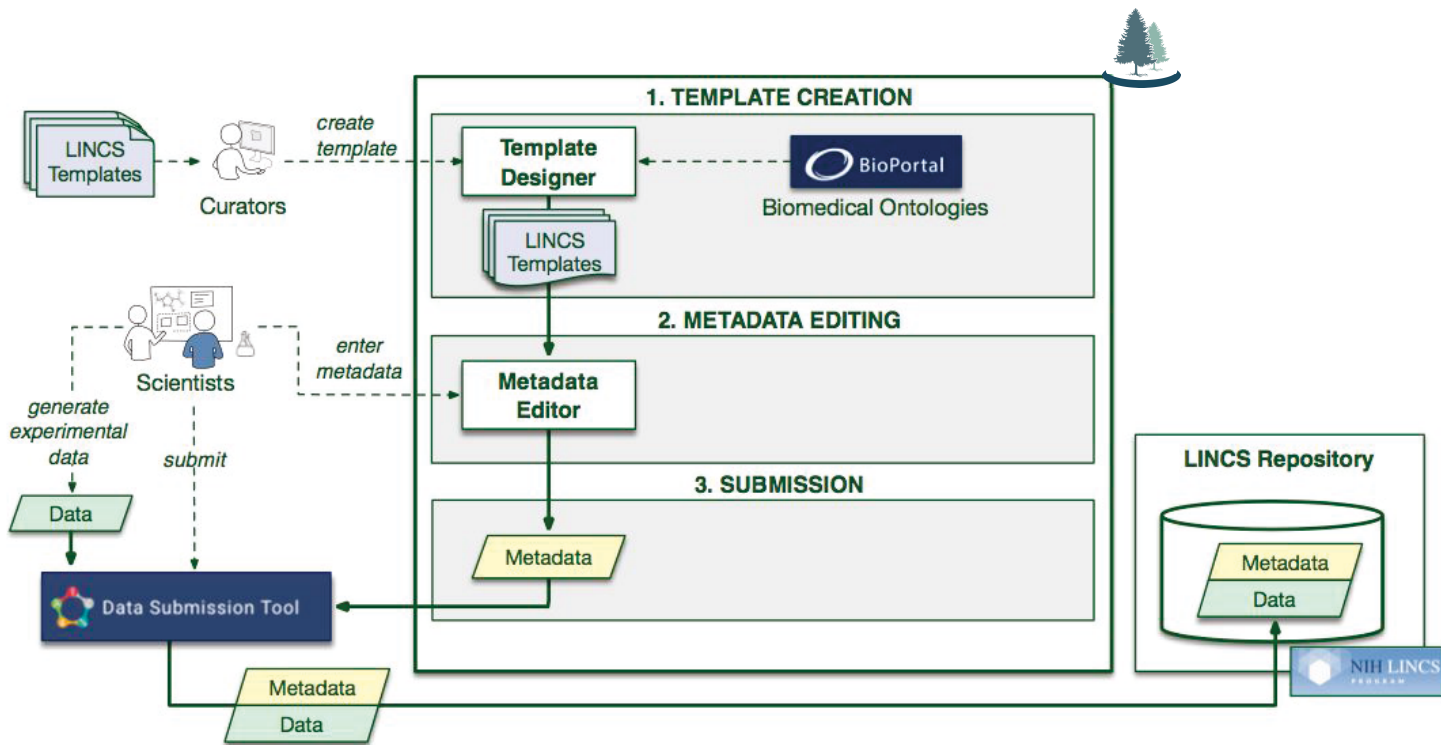
**Parameters** [i](#)

Parameter	Value	Description	Parameter Type	Data Type
q	<input type="text" value="Working"/>	Search term. It will be looked up in the resource names	query	string

# CEDAR-Only Workflow

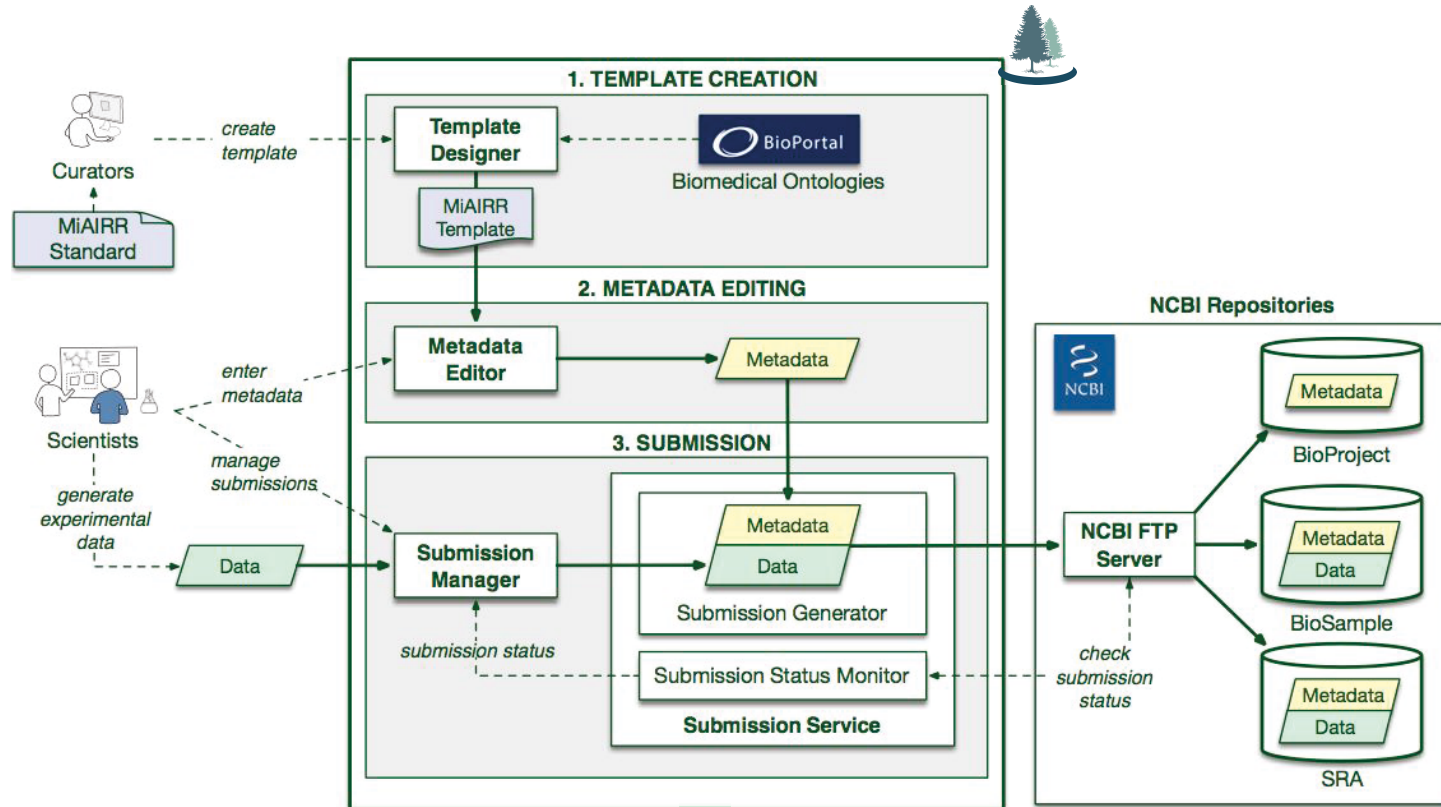


# CEDAR-LINCS Workflow



Video at <https://www.youtube.com/watch?v=5nTstseoZXg>

# CEDAR-AIRR-NCBI Workflow





New +

All / MiAIRR



### Workspace








Shared with Me

FILTER

TYPE



VERSION

Title	Crea...	Modi...	
 MiAIRR	3:35 PM	3:35 PM	 
 BioSample	3:34		
 BioProject	3:34		
 SRA	3:34		
 DN B cells in Multiple Sclerosis	3:28		

- Populate
- Open
- Share...
- Copy to...
- Move to...
- Rename...
- Publish version...
- Create version...
- Submit...
- Delete



### Share settings






#### With people



#### With groups



[Group settings !\[\]\(c50c8b7b2cc2cf9ff925edec0ee94c0d\_img.jpg\)](#)

Mark Musen	owner 	
John Graybeal	can read	
CEDAR Dev Team	can read	

**DONE**

# 3. Q&A and discussion

# References

Center for Expanded Data Annotation and Retrieval (CEDAR), <https://metadacenter.org/>.

Generic Dataset Metadata Template, <https://github.com/fair-data-collective/generic-dataset-metadata-template>.

Vasiljevic, Nikola. 2021. Generic Dataset Metadata Templates. Presentation at EOSC-Nordic workshop: FAIRification STEP 3 – generic metadata standards (<https://indico.neic.no/event/181/>). Sildes available at <https://indico.neic.no/event/181/attachments/232/389/EOSC-Nordic%20Step3%20Nikola%20Vasiljevich%2020210429.pdf>.

Wittenburg, P., H. Pergl Sustkova, A. Montesanti, S. M. Bloemers, S. H. de Waard, M. A. Musen, J. B. Graybeal, et al. 2019. “The FAIR Funder Pilot Programme to Make It Easy for Funders to Require and for Grantees to Produce FAIR Data.” ArXiv:1902.11162 [Cs], March. <http://arxiv.org/abs/1902.11162>.

# Acknowledgments

Some of the work presented in this presentation was funded by the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782.



**SSHOC**  
social sciences & humanities open cloud