Technical
University of
Denmark

BMIR

RDA P17
IG RDMinEng session, Breakout 7
Thursday, 22 April 2021 10:15 - 11:45 UTC

DOI: 10.5281/zenodo.4705970

FAIR DATA COLLECTIVE
FAIR MADE EASY

Nikola Vasiljevic and John Graybeal

with support, advocacy and early adoption by Anders Conrad, Erik Schultes and Barbara Magagna

Need(s)

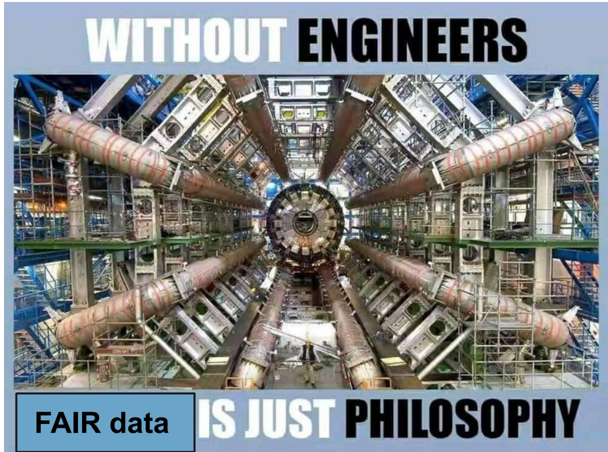
- Scalable end-to-end solution(s) for the implementation of the FAIR principles in production
    - Simple tools across the entire FAIRification lifecycle
    - Practical training targeting topics related to the creation of FAIR machine-actionable:
        - Controlled vocabularies
        - Metadata

Turning needs into solutions

- *It is 'all' about being a meticulous engineer*

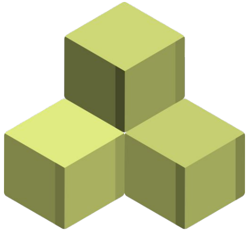
  yielding specs from needs

  creating solutions based on specs !



Controlled vocabulary specs

RDF

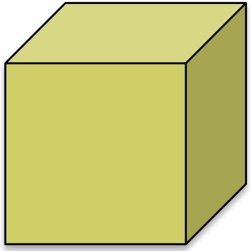


DATA
MODEL

TURTLE

JSON-LD

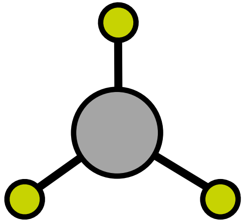
XML-RDF



FORMAT

SKOS

OWL



REPRESENTATION
LANGUAGE

Why RDF, Turtle and SKOS?

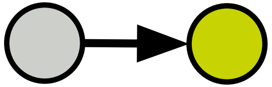
- **RDF** (Resource Data Framework) is a standard model for information (e.g. vocabularies) interchange on the Web

- **Turtle** is a common, human-readable and very compact data format for storing RDF data

- **SKOS** (Simple Knowledge Organization System) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or **any other type of structured controlled vocabulary**.

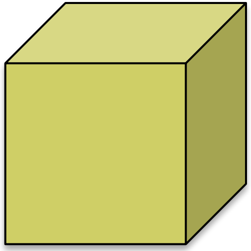
Metadata specs

LINKED DATA



APPROACH

JSON-LD
TURTLE
XML-RDF



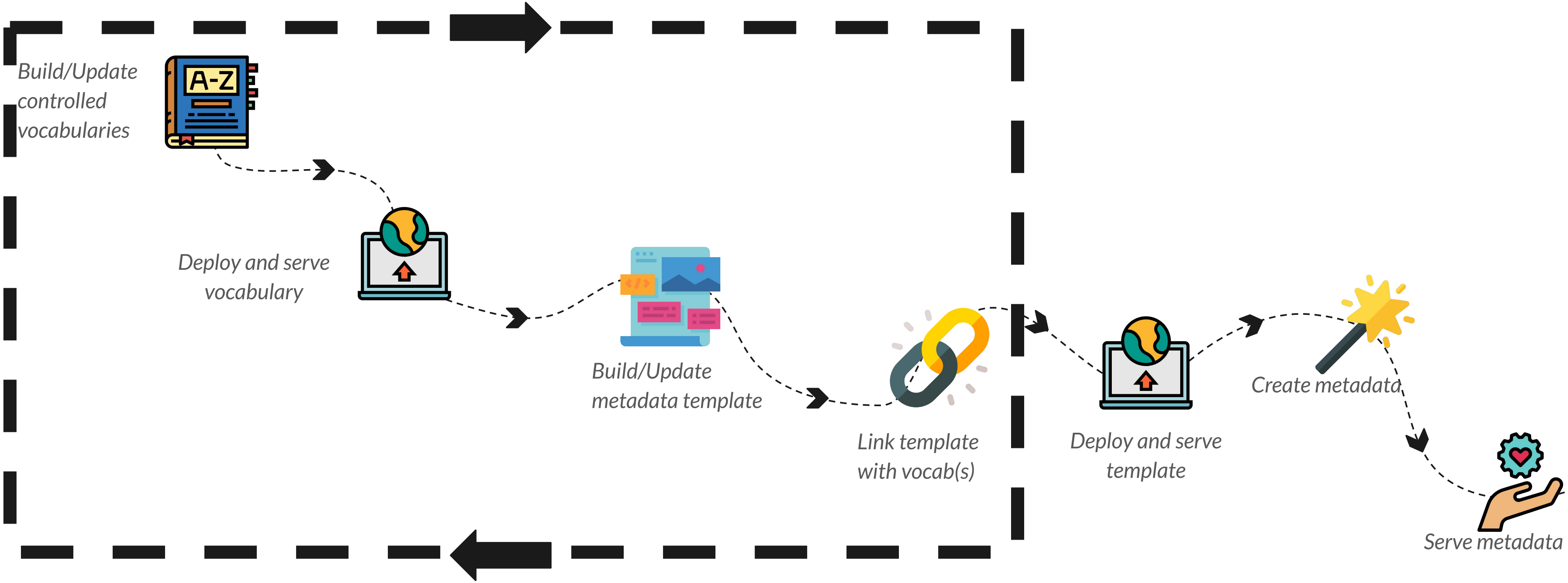
FORMAT

Why LINKED DATA and JSON-LD?

- **LINKED DATA** builds upon standard Web technologies such as HTTP and URIs/IRIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by machines. This enables data from different sources to be connected and queried.

- **JSON-LD** is a lightweight Linked Data format. It is easy for humans to read and write. It is based on the already successful JSON format and provides a way to help JSON data interoperate at Web-scale. JSON-LD is an ideal data format for programming environments, REST Web services, and unstructured databases such as Apache CouchDB and MongoDB.

FAIRification roadmap


Build/Update controlled vocabularies
A-Z
Deploy and serve vocabulary
Build/Update metadata template
Link template with vocab(s)
Deploy and serve template
Create metadata
Serve metadata

Icons made by https://www.freepik.com

Solutions

Template

- **Generic Dataset Metadata Template (GDMT)** - domain agnostic machine-actionable metadata template
- Google Sheet / Excel template for machine-actionable controlled vocabulary creation (see sheet2rdf and excel2rdf)

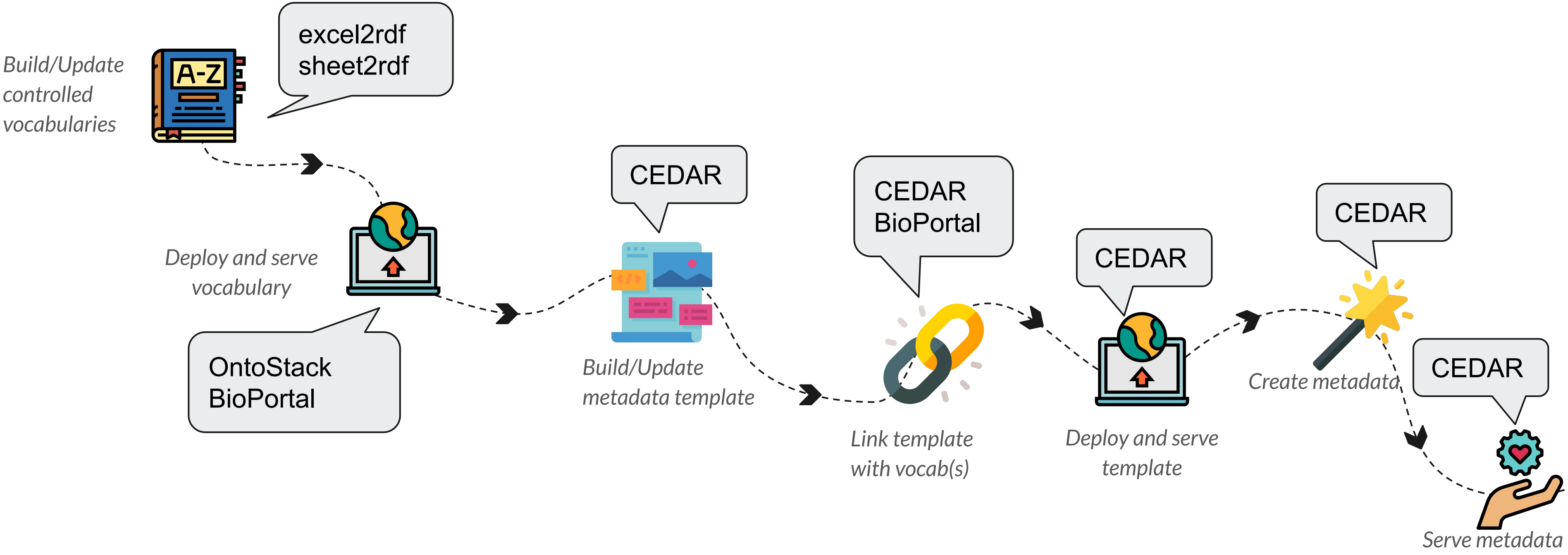
Tools

- excel2rdf / sheet2rdf - automatic workflows for building machine-actionable controlled vocabularies using Excel / Google Sheets
- OntoStack - graph database, SPARQL endpoint, URI resolver and ontology browser all in one
- CEDAR WorkBench - online collaborative platform for creation of metadata templates and their instances (i.e., metadata)
- BioPortal - ontology registration and indexing service

Training

- **Rapid Metadata 4 Machine (M4M) Workshops** - low-barrier course that train participants to build controlled vocabularies and metadata templates and create machine-actionable metadata using the above tools.

FAIRification roadmap



Build/Update controlled vocabularies

excel2rdf
sheet2rdf

Deploy and serve vocabulary

OntoStack
BioPortal

Build/Update metadata template

CEDAR

Link template with vocab(s)

CEDAR
BioPortal

Deploy and serve template

CEDAR

Create metadata

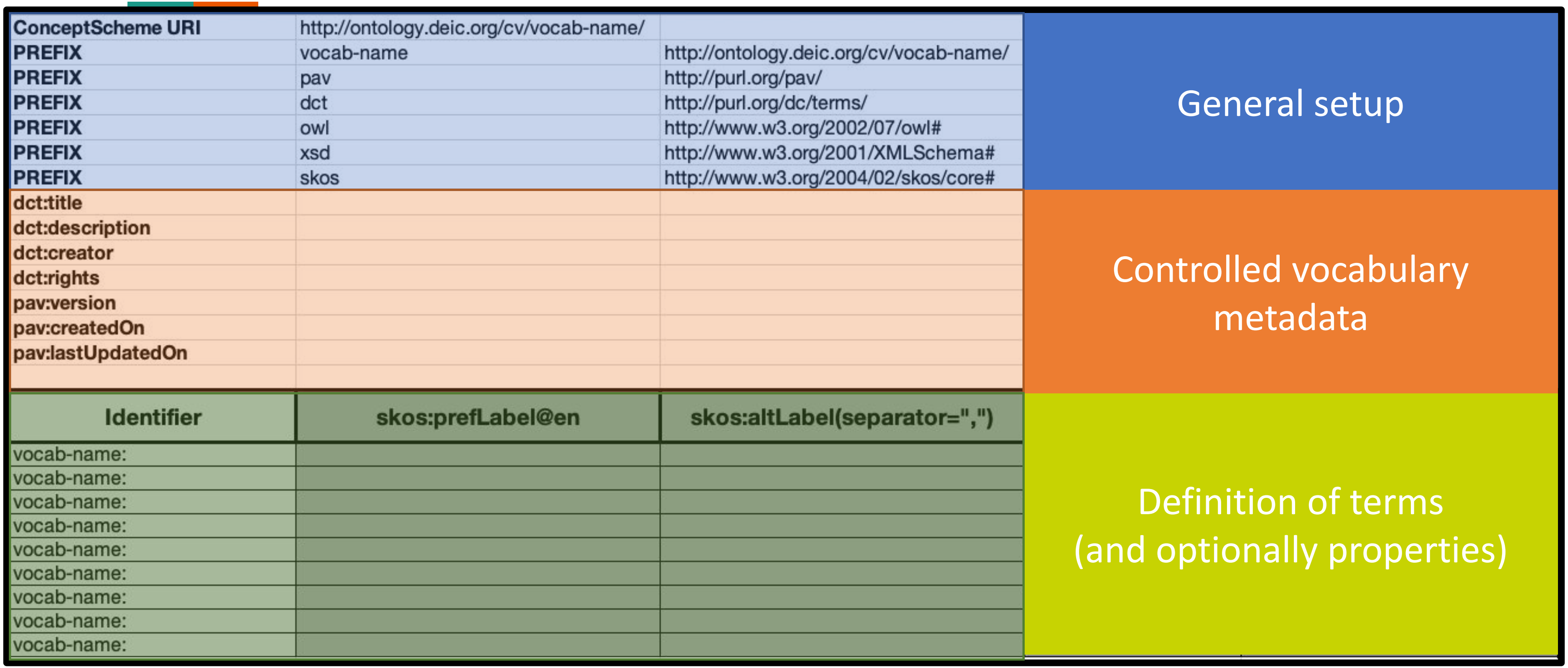
CEDAR

Serve metadata

CEDAR

Icons made by https://www.freepik.com

Controlled vocab template based on SKOS Play!



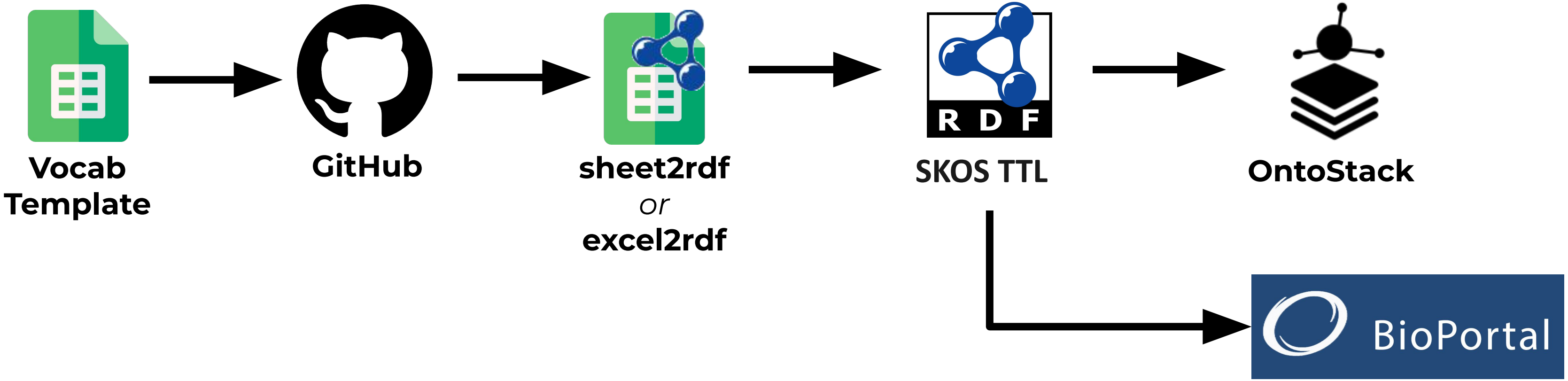
ConceptScheme URI	http://ontology.deic.org/cv/vocab-name/	
PREFIX	vocab-name	http://ontology.deic.org/cv/vocab-name/
PREFIX	pav	http://purl.org/pav/
PREFIX	dct	http://purl.org/dc/terms/
PREFIX	owl	http://www.w3.org/2002/07/owl#
PREFIX	xsd	http://www.w3.org/2001/XMLSchema#
PREFIX	skos	http://www.w3.org/2004/02/skos/core#
dct:title		
dct:description		
dct:creator		
dct:rights		
pav:version		
pav:createdOn		
pav:lastUpdatedOn		
Identifier	**skos:prefLabel@en**	**skos:altLabel(separator=",")**
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		
vocab-name:		

General setup

Controlled vocabulary metadata

Definition of terms
(and optionally properties)

excel2rdf & sheet2rdf



Icons made by https://www.freepik.com

Excel2RDF and Sheet2RDF

- Automatic workflows executed by means of GitHub actions

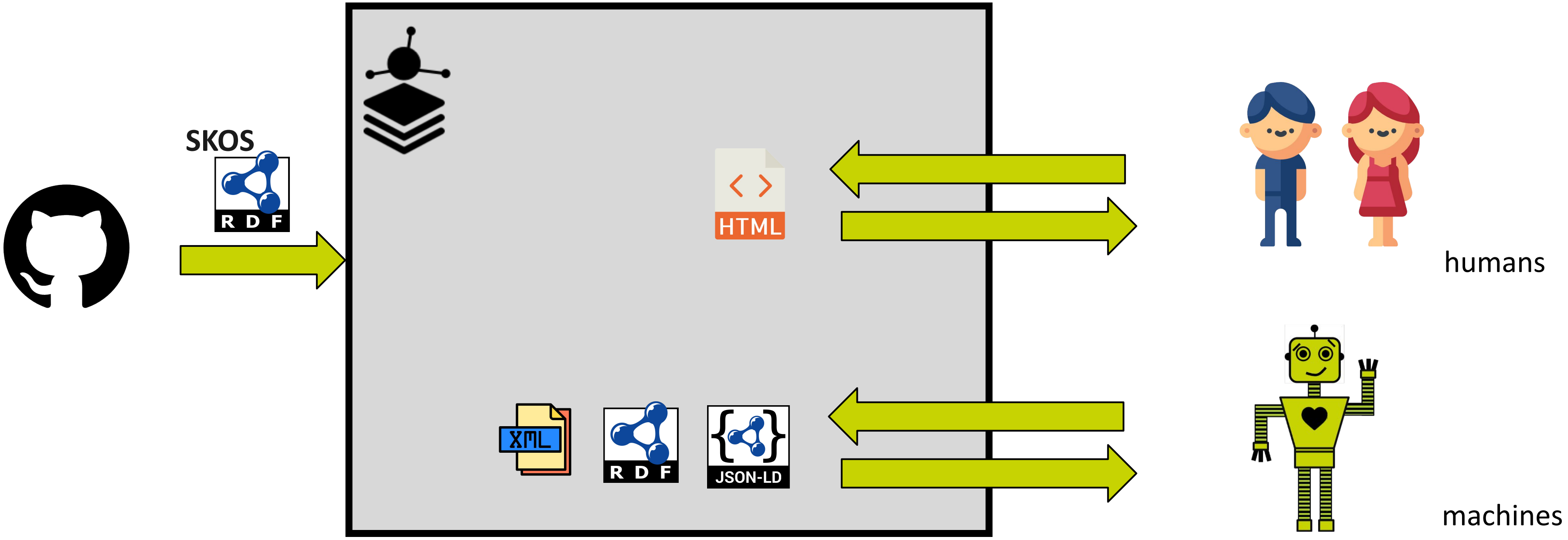
- Contains underlying shell, python and java programs which:
  - (1) converts Excel/Google Sheet to the machine-actionable controlled vocabulary using xls2rdf
  - (2) tests the derived controlled vocabulary using qSKOS
  - (3) commits the conversion results and tests logs to a Git repository
  - (4) deploys the vocabulary to **OntoStack to be served to humans and machines**

- Currently an initial manual configuration is required to set up BioPortal to fetch and index vocabulary

- **Workflows** are used by: VODAN Africa and Asia, ZonMW Covid program, DTU Wind Energy, DeiC, International Energy Agency WIND Task 32

OntoStack

A set of orchestrated micro-services configured and interfaced such that they can intake terminologies and serve them to humans or machines



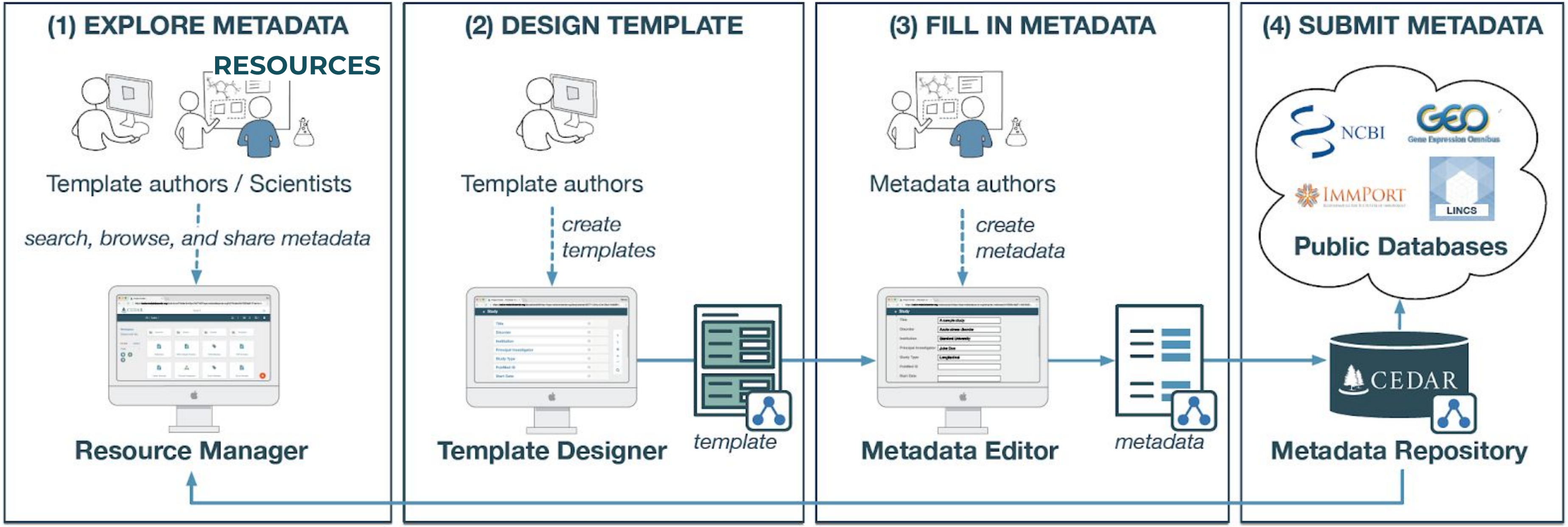
OntoStack in its core

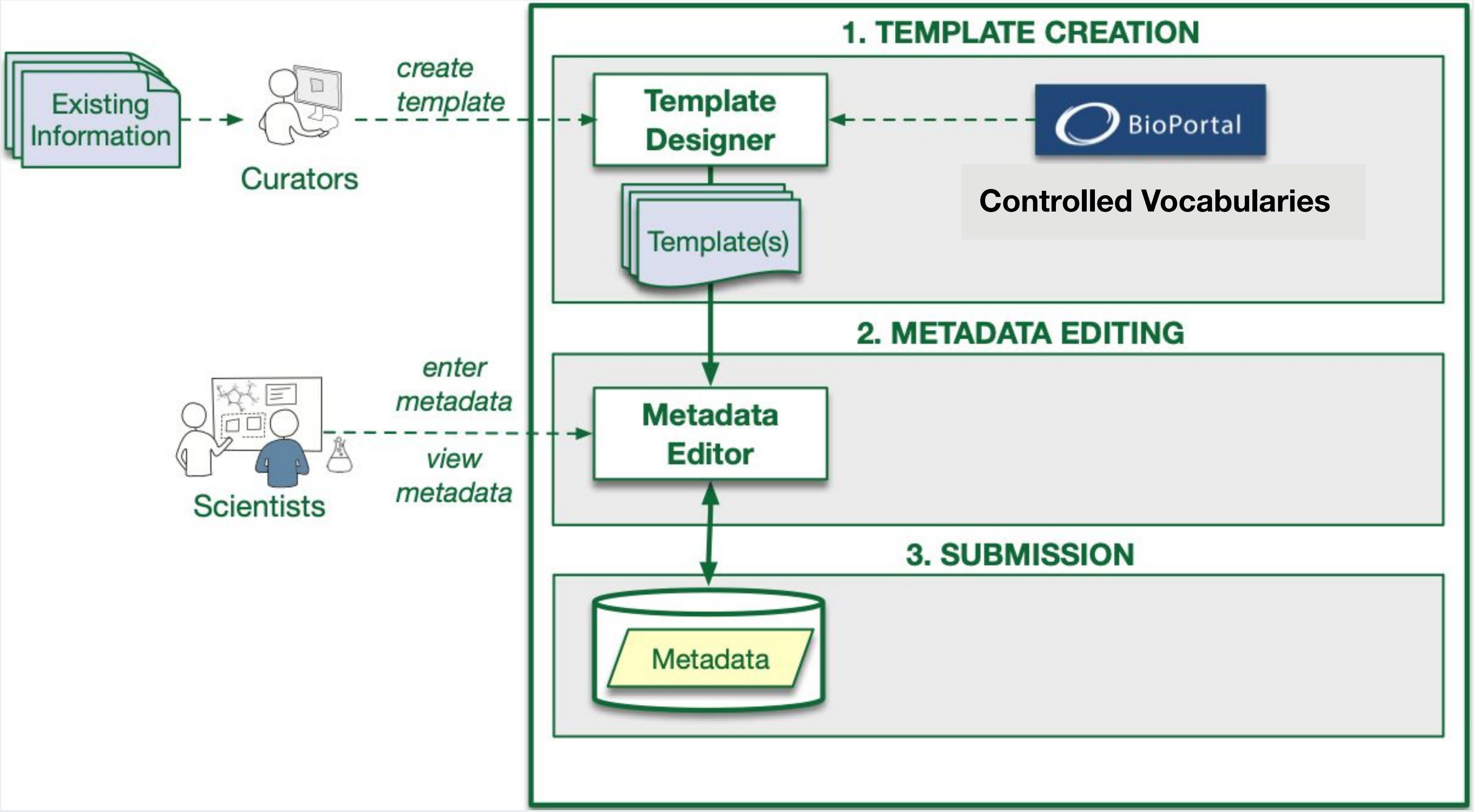
- A set of orchestrated micro-services:
  - Edge router/URI resolver (**Traefik**)
  - Graph database (**Apache Jena Fuseki**)
  - Web-based terminology browser/UI (**SKOSMOS**)
- Four instances of OntoStack:
  - Departmental: http://data.windenergy.dtu.dk/ontologies/view
  - National: http://ontology.deic.dk
  - International: http://vocab.fairdatacollective.org/ and http://vocab.ieawindtask32.org/

BioPortal

- Upload/register ontology to its database
- Browse the library of ontologies
- Search for a term across multiple ontologies
- Browse mappings between terms in different ontologies
- Receive recommendations on which ontologies are most relevant for a corpus
- Annotate text with terms from ontologies
- …
- **Make ontology visible/accessible in CEDAR Workbench**

CEDAR WorkBench





Controlled Vocabularies

Generic Dataset Metadata Template (GDMT)

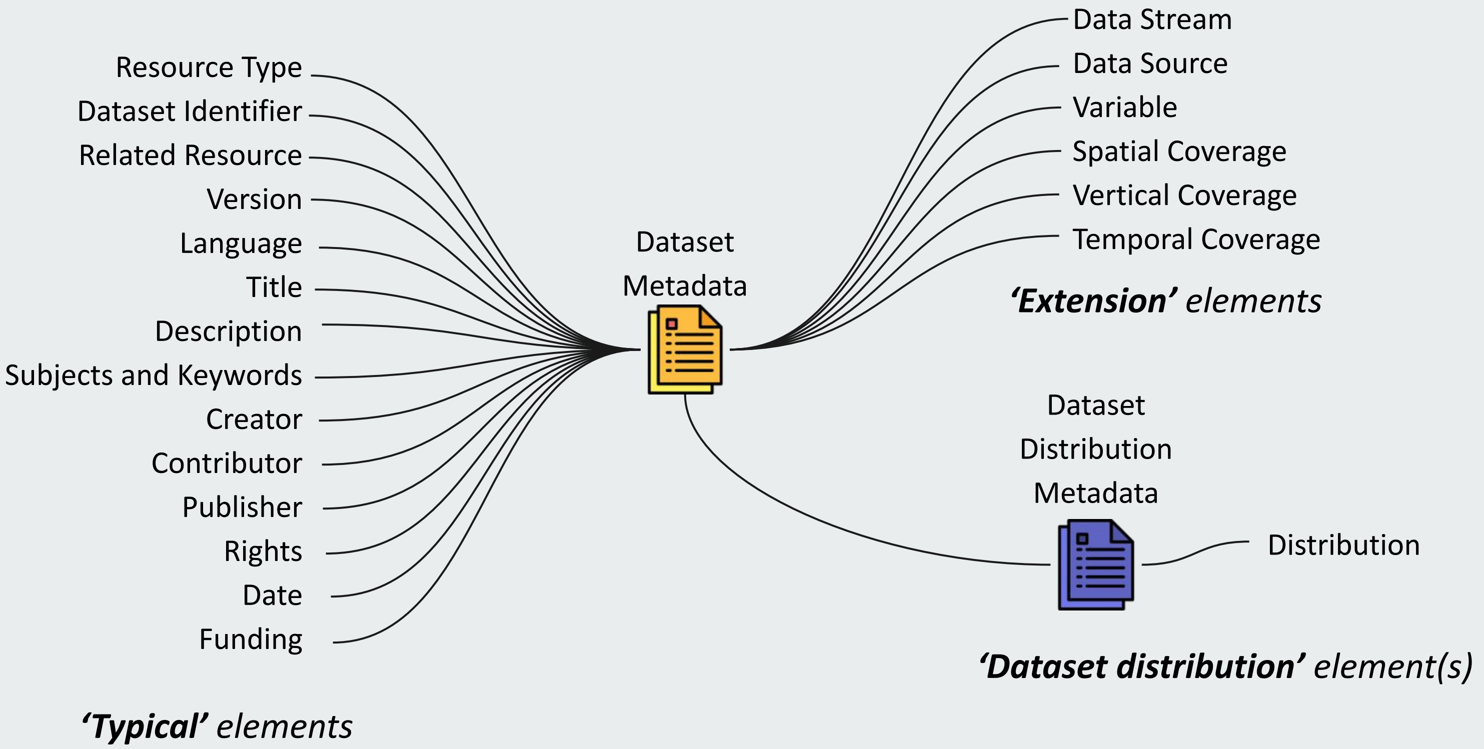
- Inspired by DataCite and DCAT scheme
- Scheme fused, improved, extended and ‘simplified’

- GDMT contains **100 fields** (‘only’ **13** mandatory) grouped in **~20 elements**

- Unlike the DataCite template, GDMT is MACHINE-ACTIONABLE, details at:
  - CEDAR
  - GitHub

- GDMT contains a **‘back-end’ vocabulary** that enables machine-actionability, which contains:
  - ~**130** RDF properties
  - ~**1000** controlled terms
- The development of the template partially funded by DeiC and ZonMW Covid program, but largely by our free time

FAIR
Data Collective



Resource Type
Dataset Identifier
Related Resource
Version
Language
Title
Description
Subjects and Keywords
Creator
Contributor
Publisher
Rights
Date
Funding

*'**Typical**' elements*

Dataset Metadata

Data Stream
Data Source
Variable
Spatial Coverage
Vertical Coverage
Temporal Coverage

*'**Extension**' elements*

Dataset Distribution Metadata

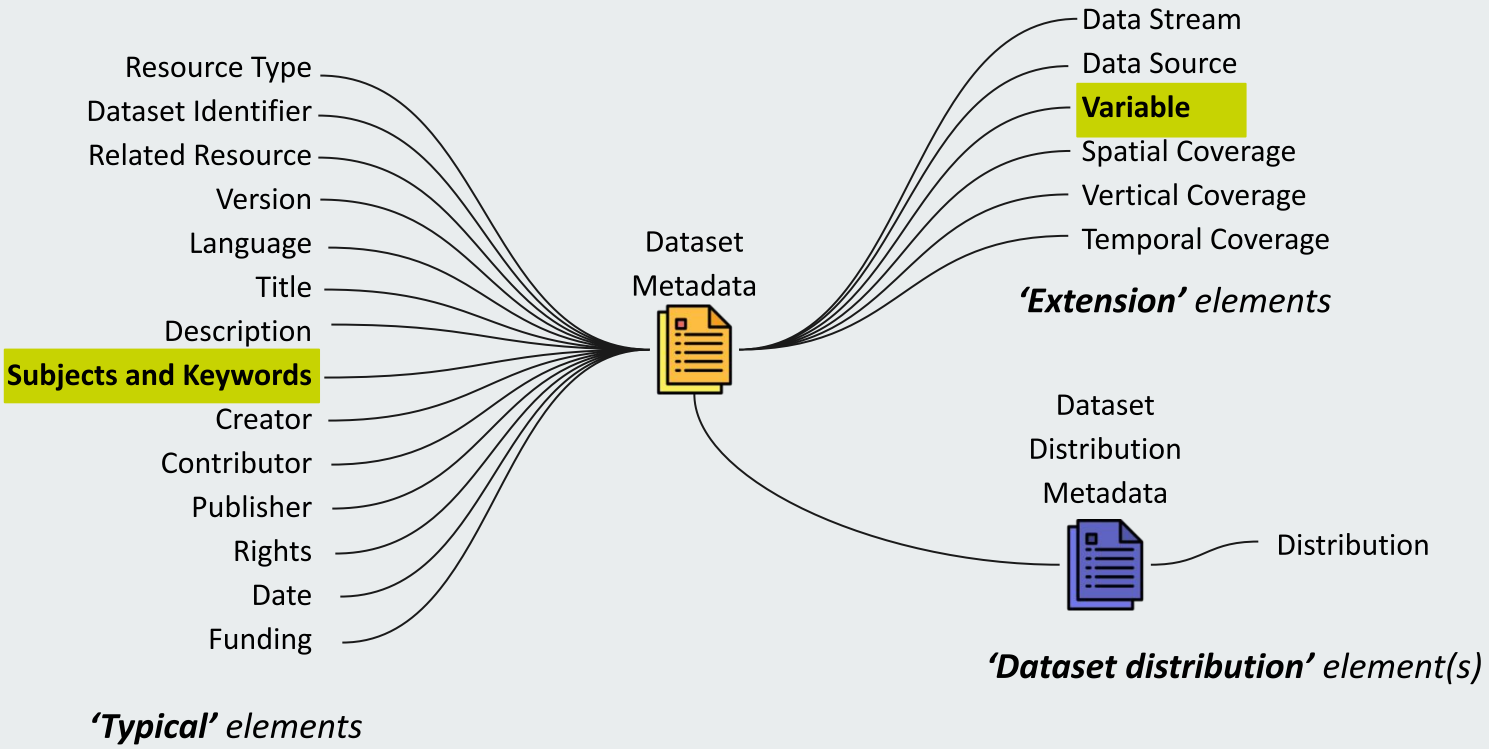
Distribution

*'**Dataset distribution**' element(s)*

You can find definitions of elements and fields on CEDAR and GitHub.

OntoStack serves the GDMT ontology, which contains a number of controlled terms and RDF properties that enable machine-actionability .

FAIR
Data Collective



Resource Type

Dataset Identifier

Related Resource

Version

Language

Title

Description

Subjects and Keywords

Creator

Contributor

Publisher

Rights

Date

Funding

*‘**Typical**’ elements*

Dataset Metadata

Data Stream

Data Source

Variable

Spatial Coverage

Vertical Coverage

Temporal Coverage

*‘**Extension**’ elements*

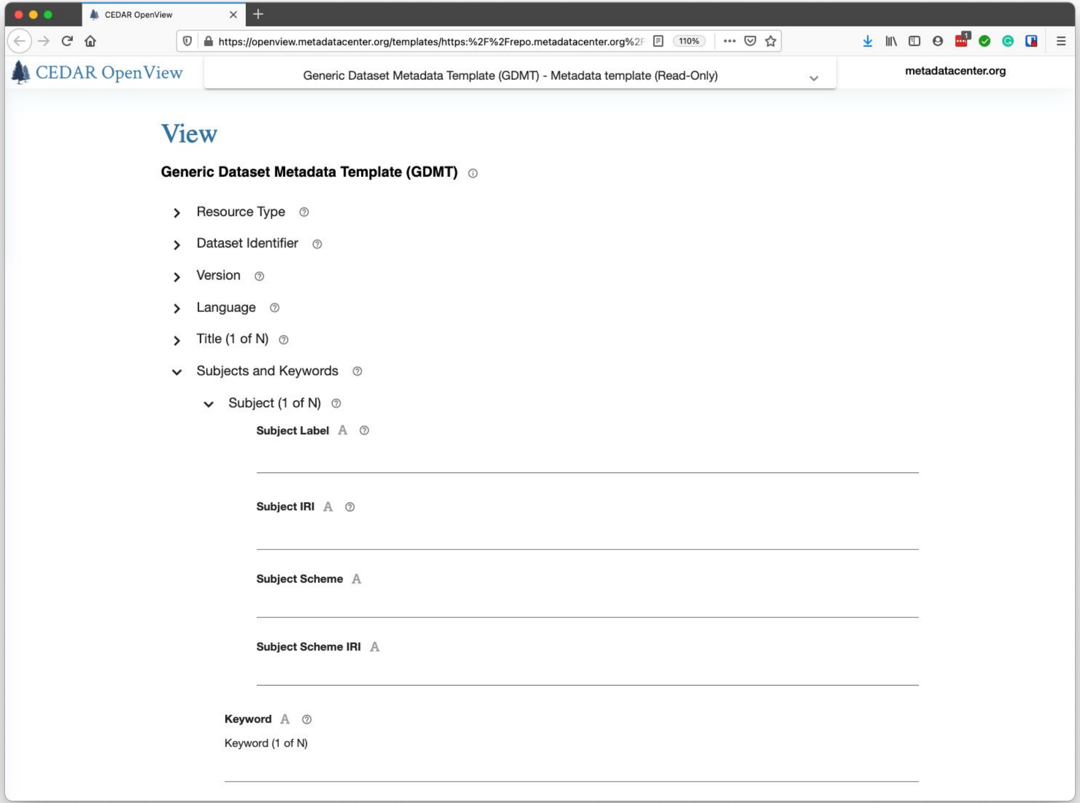
Dataset Distribution Metadata

Distribution

*‘**Dataset distribution**’ element(s)*

🟩 *By creating domain specific controlled vocabularies and updating GDMT to use them, we turn this template to be domain specific*

GDMT in CEDAR OpenView


CEDAR OpenView
https://openview.metadatacenter.org/templates/https:%2F%2Frepo.metadatacenter.org%2F
110%
metadatacenter.org

CEDAR OpenView

Generic Dataset Metadata Template (GDMT) - Metadata template (Read-Only)

View

Generic Dataset Metadata Template (GDMT)

- Resource Type
- Dataset Identifier
- Version
- Language
- Title (1 of N)
- Subjects and Keywords
  - Subject (1 of N)
    - **Subject Label**
    - **Subject IRI**
    - **Subject Scheme**
    - **Subject Scheme IRI**
  - **Keyword**
    - Keyword (1 of N)

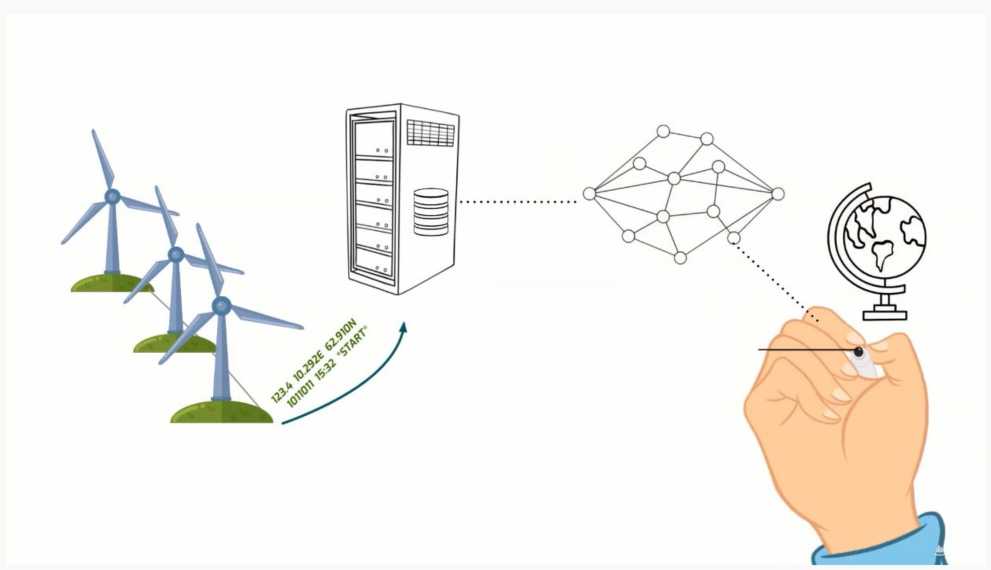
Rapid M4M* Workshops

- Low-barrier training
- Instructing participants in :
  - How to use tooling presented in previous slides
  - How to build machine actionable controlled vocabularies and metadata templates
  - Create machine-actionable metadata
- Rapid M4Ms take place over 2 x 0.5 days
- Almost exclusively focused on practical work instead of theory
- Development of the rapid M4M training material funded by DeiC
- This training is offered in Denmark by DeiC, while globally the training is organized via Go FAIR Foundation

Visual training material with simple tools = simplifying FAIR data principles implementation

*Read more about the origin of M4M workshops here: https://www.go-fair.org/today/making-fair-metadata/

See Wind Energy use case

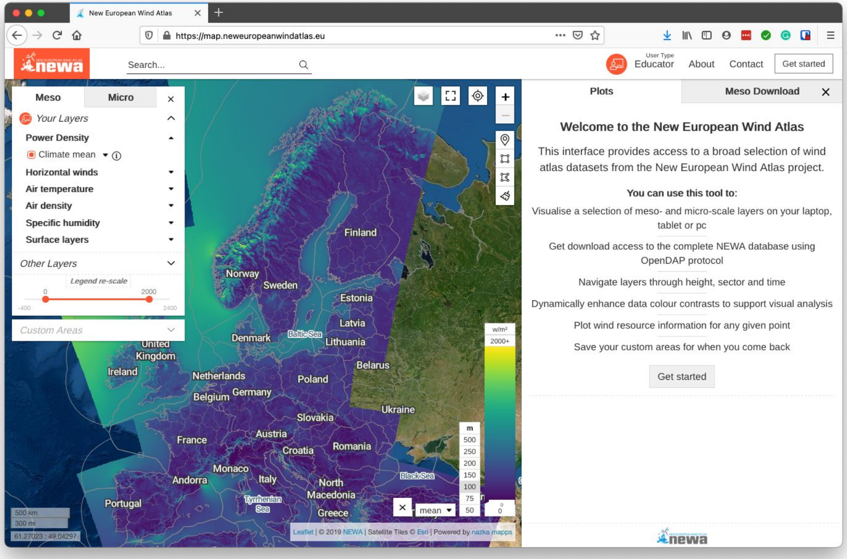


http://bit.ly/we-fair

Future work

- Automate metadata generation, i.e. reduce or completely remove a need for human interaction

- The idea will be implemented as one of the features of:
  RESTful API for [New European Wind Atlas](https://map.neweuropeanwindatlas.eu) micro scale data subsetting and aggregation



Welcome to the New European Wind Atlas

This interface provides access to a broad selection of wind atlas datasets from the New European Wind Atlas project.

You can use this tool to:

Visualise a selection of meso- and micro-scale layers on your laptop, tablet or pc

Get download access to the complete NEWA database using OpenDAP protocol

Navigate layers through height, sector and time

Dynamically enhance data colour contrasts to support visual analysis

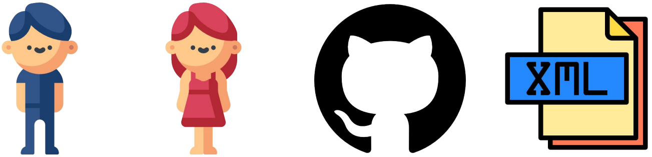
Plot wind resource information for any given point

Save your custom areas for when you come back

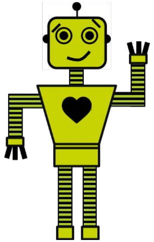
Get started

Thank you.

Source of material

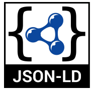


Icons made by https://www.freepik.com



Vasiljevic, Nikola. (2021). **MetaManMachine**. Zenodo. http://doi.org/10.5281/zenodo.4471098
Licensed under: CC BY-SA 4.0

Logo made by Bill Schwappacher <bill@tracermedia.com>
provided by W3C for public use



Released by https://json-ld.org/ under CC0