DOI: 10.5281/zenodo.4726001

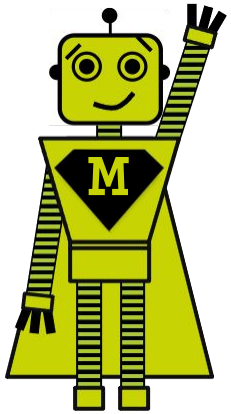
Generic Dataset Metadata Template

Nikola Vasiljevic, Technical University of Denmark - Department for Wind Energy

This presentation represents the mix of the following material:
http://doi.org/10.5281/zenodo.4621141
http://doi.org/10.5281/zenodo.4705970

This presentation contains personal opinions which might not reflect standpoints of Technical University of Denmark neither FAIR DATA COLLECTIVE.

Our protagonists

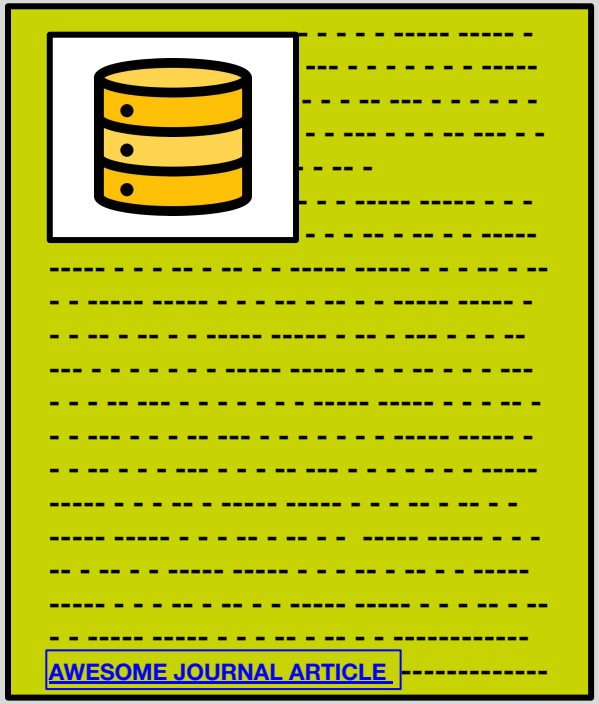


MetaManMachine (aka 3M)

Robert

Ana

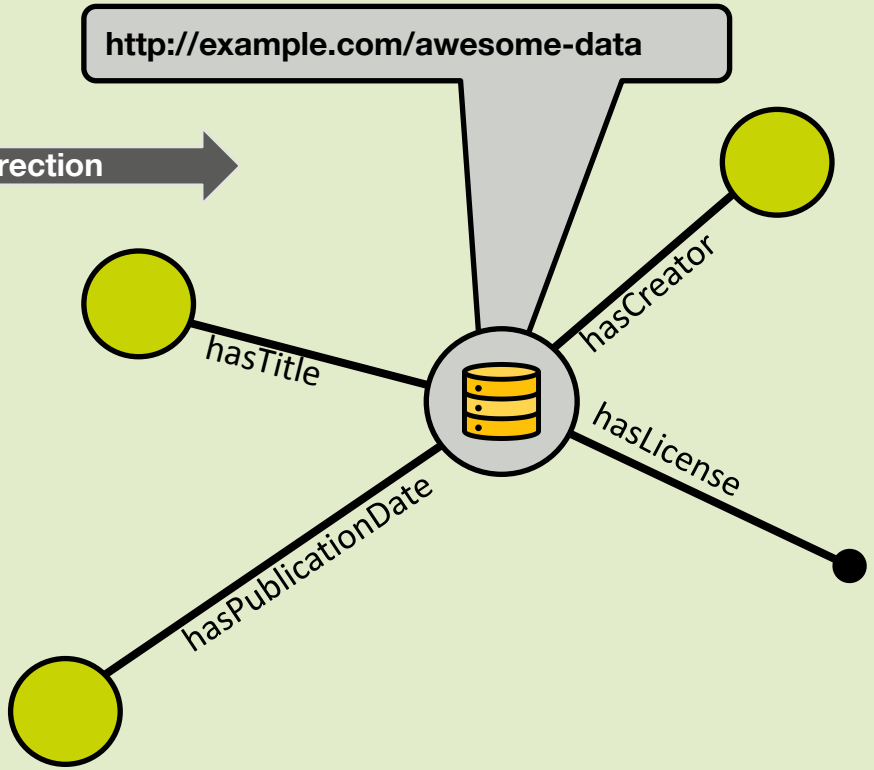
http://example.com/awesome-data

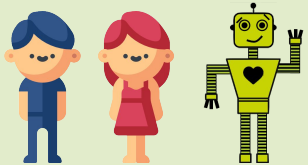


Understandable only by humans

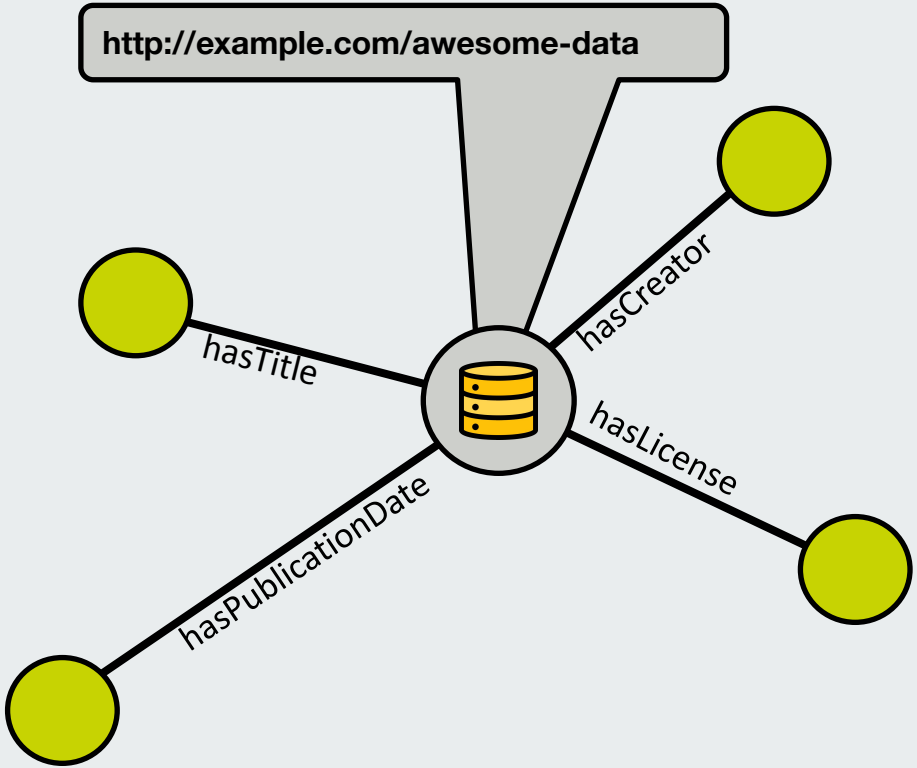
Desired direction

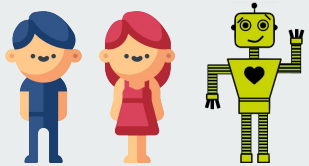
http://example.com/awesome-data



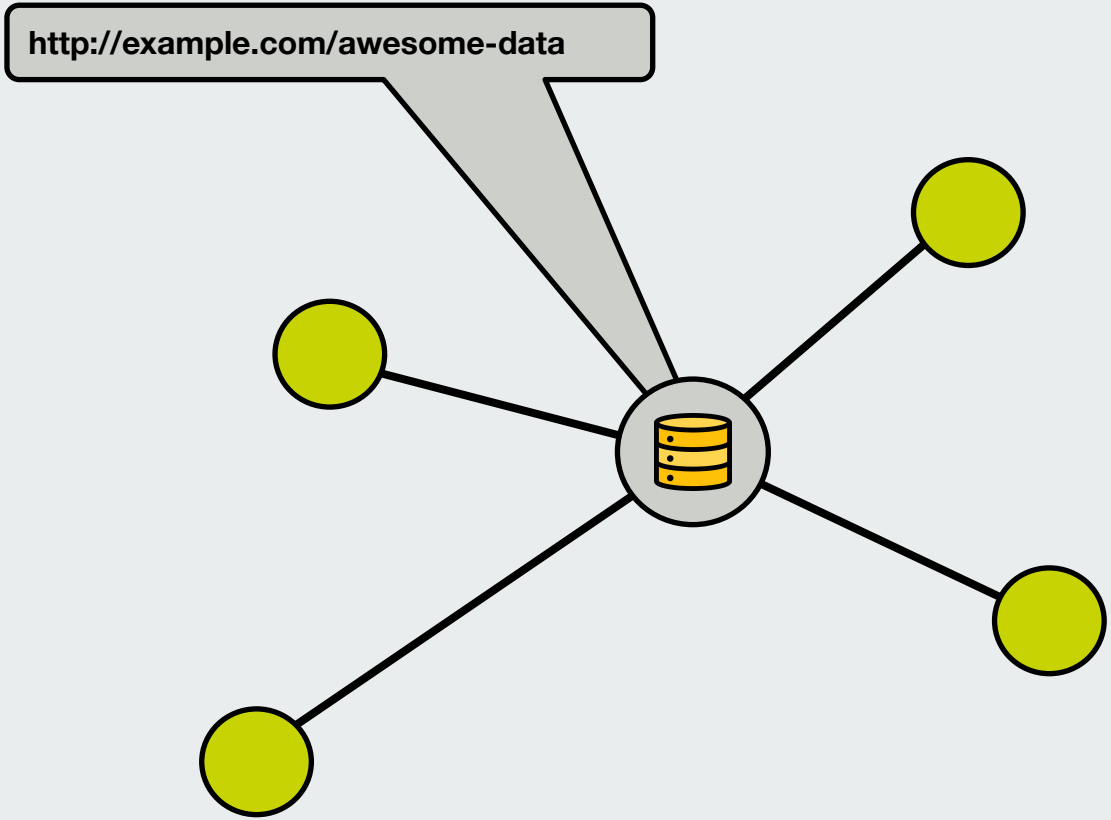


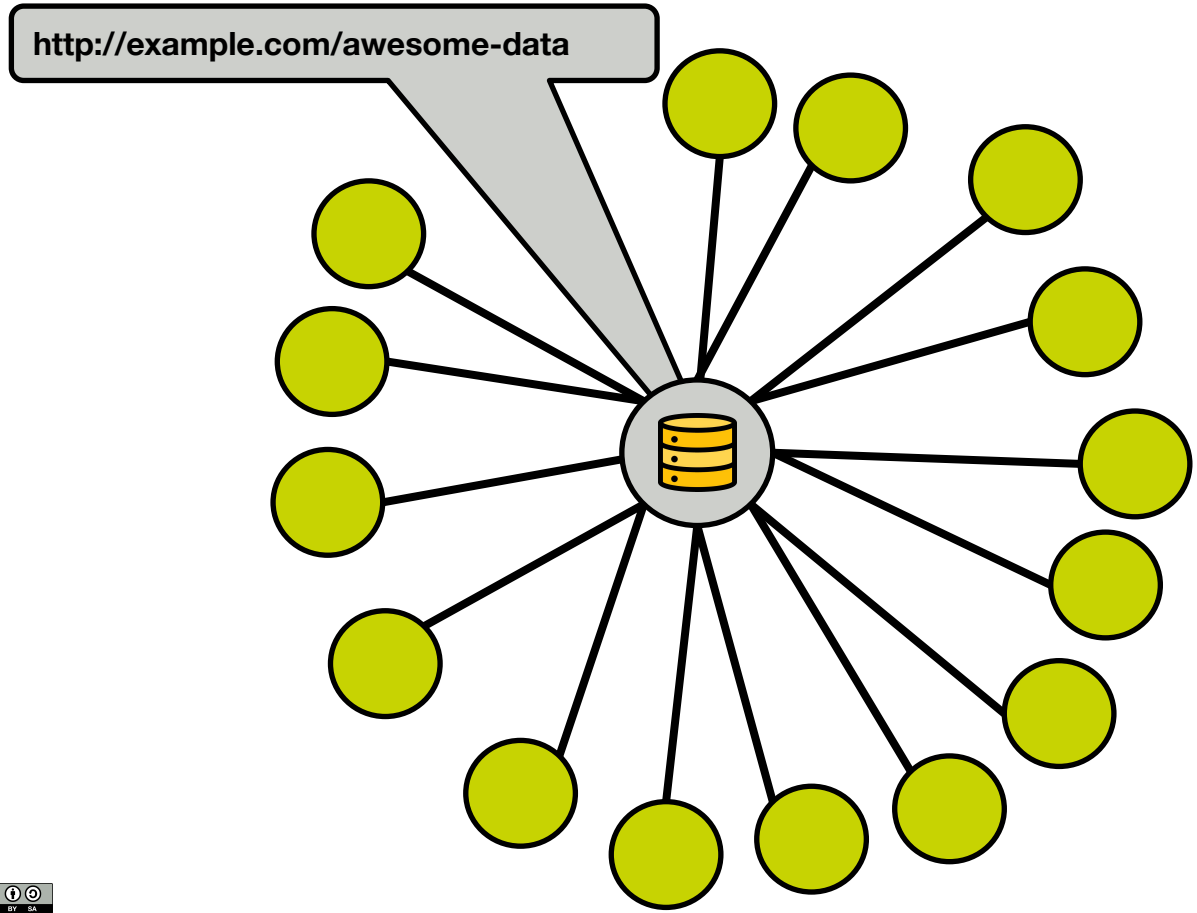
Understandable by humans and machines

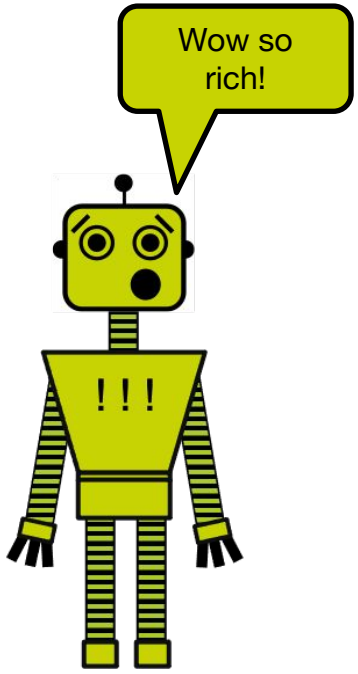




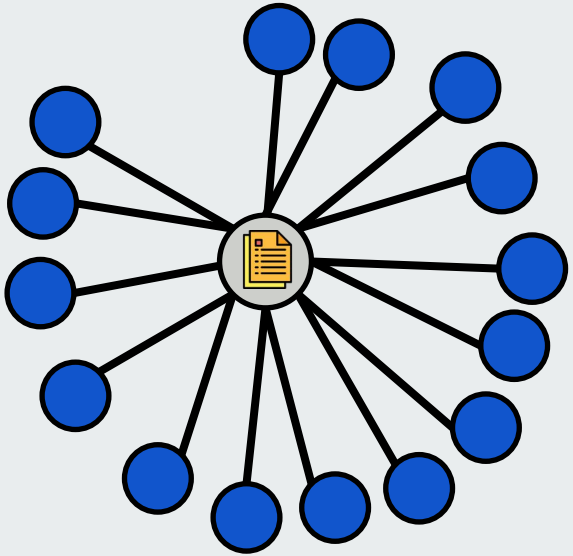
Understandable by humans and machines


http://example.com/awesome-data





Metadata Template



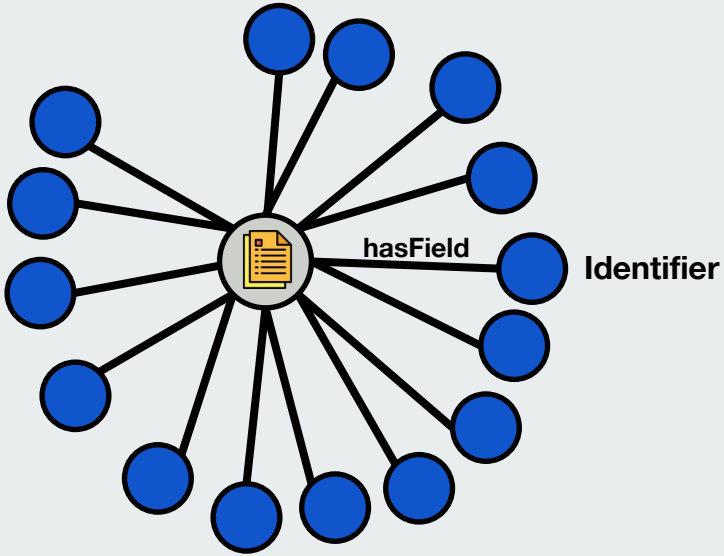
Graph representation

Field*	Value
Identifier	*insert identifier*
Title	*insert title*
Creator	*insert creator*
Publication date	*insert publication_date*
License	*insert license*
...	...

(Web) Form representation

***Field = Property**

Metadata Template



Graph representation

Field*	Value
Identifier	*insert identifier*
Title	*insert title*
Creator	*insert creator*
Publication date	*insert publication_date*
License	*insert license*
...	...

(Web) Form representation

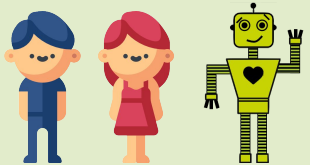
*Field = Property

Transition to machine actionable template

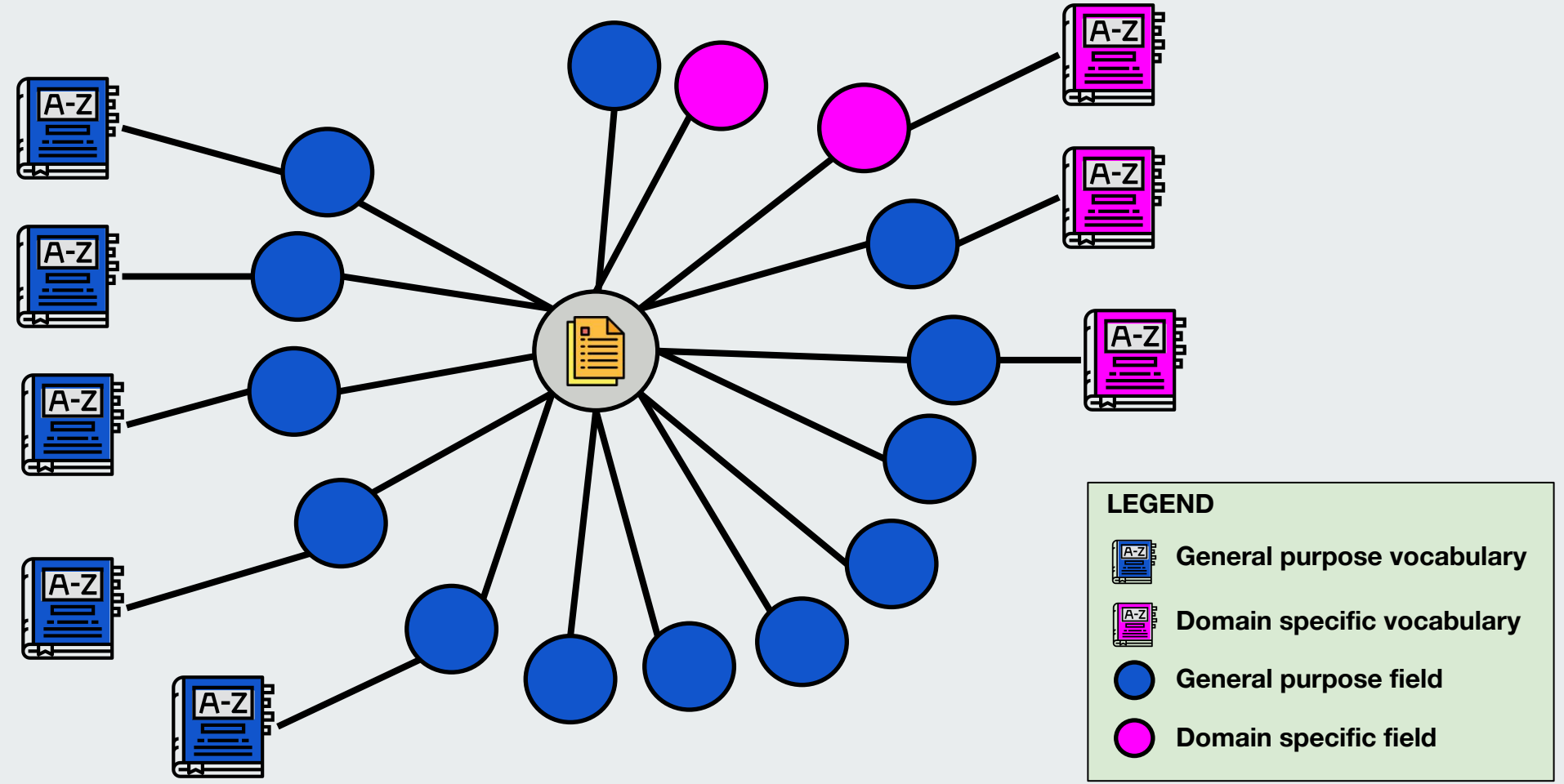
Field	Value
Title	*insert_title*
Creator	*insert_creator*
Publication date	*insert_publication_date*
License	*insert_license*
Subject	*insert_subject*
Variable	*insert_variable*
...	...

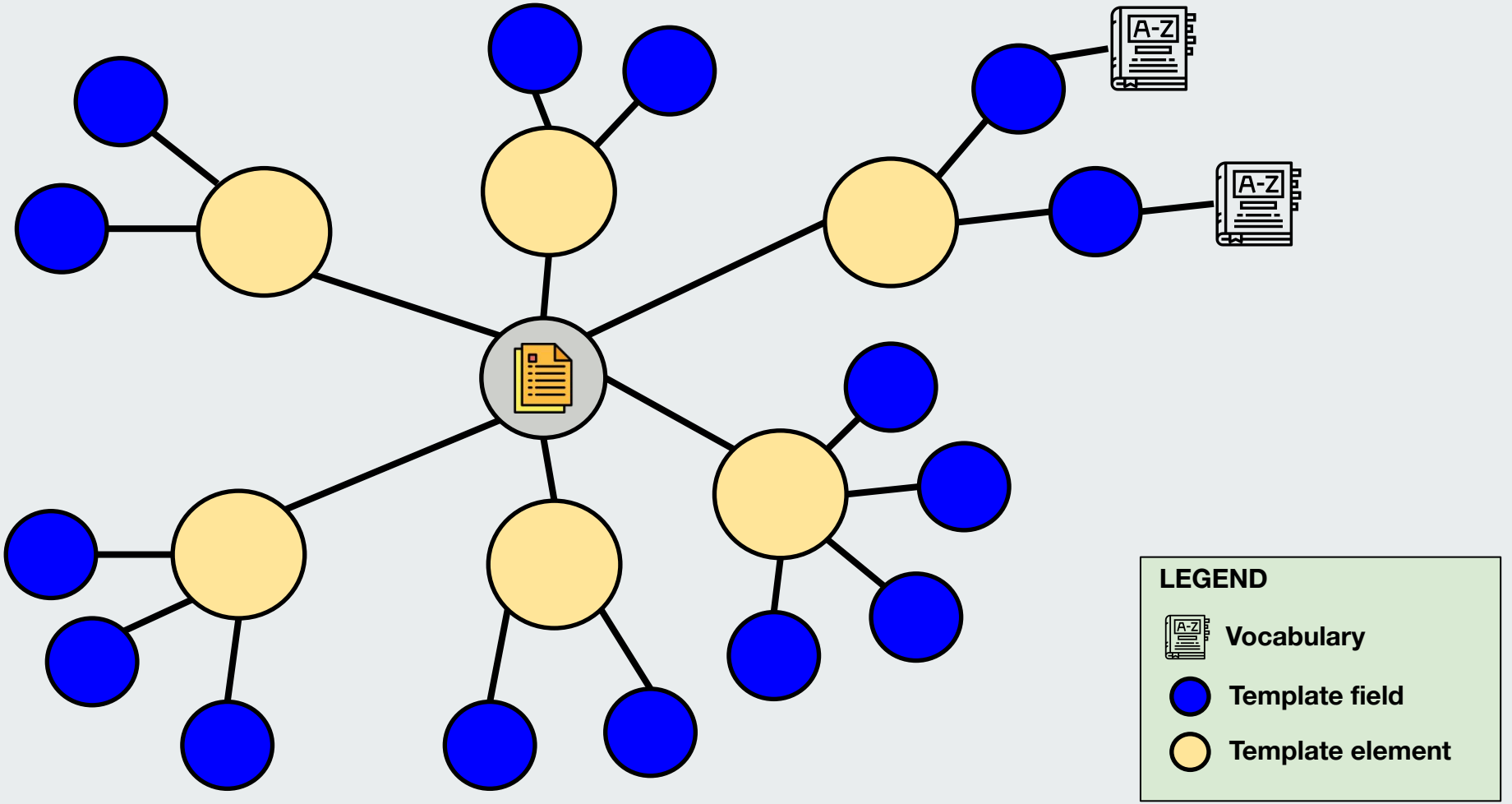
Human readable template

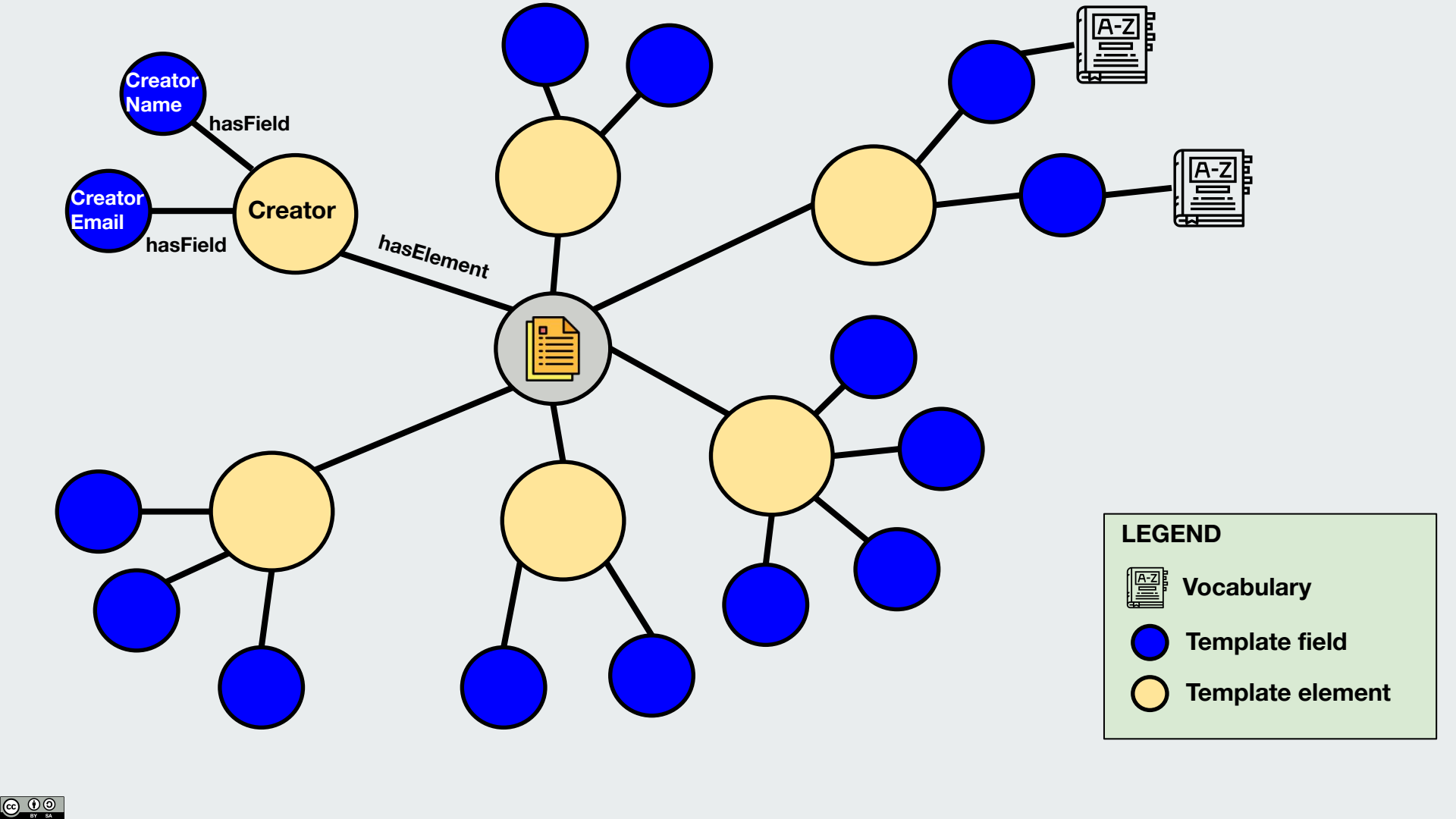
Field	Value
http://purl.org/dc/elements/1.1/title	Free text
http://purl.org/dc/elements/1.1/creator	URL representing ORCID ID
http://vocab.fairdatacollective.org/gdmt/hasDatasetDate	datetime string
http://purl.org/dc/elements/1.1/rights	**https://spdx.org/licenses/**
http://purl.org/dc/elements/1.1/subject	**http://data.windenergy.dtu.dk/controlled-terminology/taxonomy-topics/**
http://vocab.fairdatacollective.org/gdmt/hasVariableInfo	**http://data.windenergy.dtu.dk/controlled-terminology/wind-energy-parameters/**
...	...



Human and machine actionable template

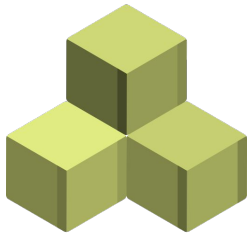

A-Z
A-Z
A-Z
A-Z
A-Z
A-Z
A-Z
A-Z
LEGEND
General purpose vocabulary
Domain specific vocabulary
General purpose field
Domain specific field


A-Z
A-Z
LEGEND
Vocabulary
Template field
Template element


Creator Name
hasField
Creator Email
hasField
Creator
hasElement
A-Z
A-Z
LEGEND
Vocabulary
Template field
Template element

Controlled vocabulary specs

RDF

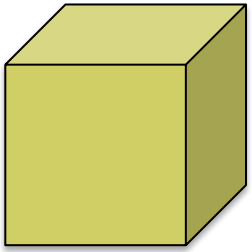


DATA MODEL

TURTLE

JSON-LD

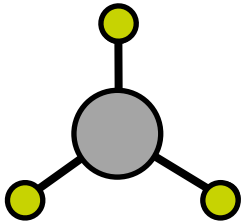
XML-RDF



FORMAT

SKOS

OWL



REPRESENTATION LANGUAGE

Why RDF, Turtle and SKOS?

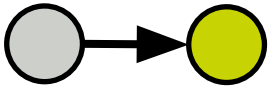
- **RDF** (Resource Data Framework) is a standard model for information (e.g. vocabularies) interchange on the Web

- **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**.

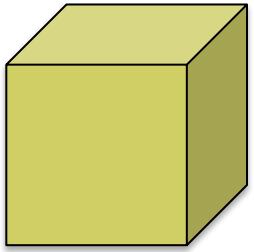
Metadata specs

LINKED DATA



APPROACH

JSON-LD

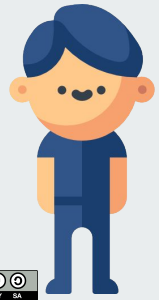


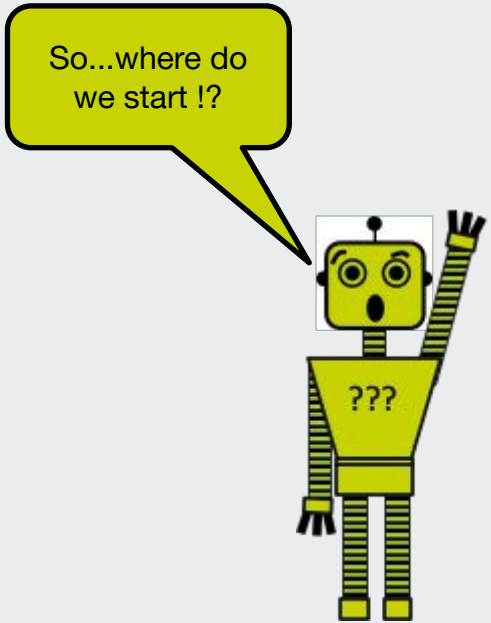
FORMAT

Why LINKED DATA and JSON-LD?

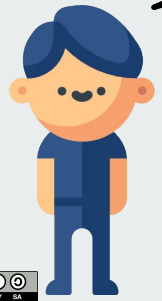
- **LINKED DATA** builds upon standard Web technologies such as HTTP and URIs/IRIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by machines. This enables data from different sources to be connected and queried.

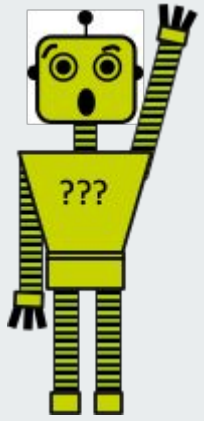
- **JSON-LD** is a lightweight Linked Data format. It is easy for humans to read and write. It is based on the already successful JSON format and provides a way to help JSON data interoperate at Web-scale. **JSON-LD is an ideal data format for programming environments**, REST Web services, and unstructured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





We start with a short intro to the Generic Dataset Metadata Template (**GDMT**)





Generic Dataset Metadata Template (GDMT)

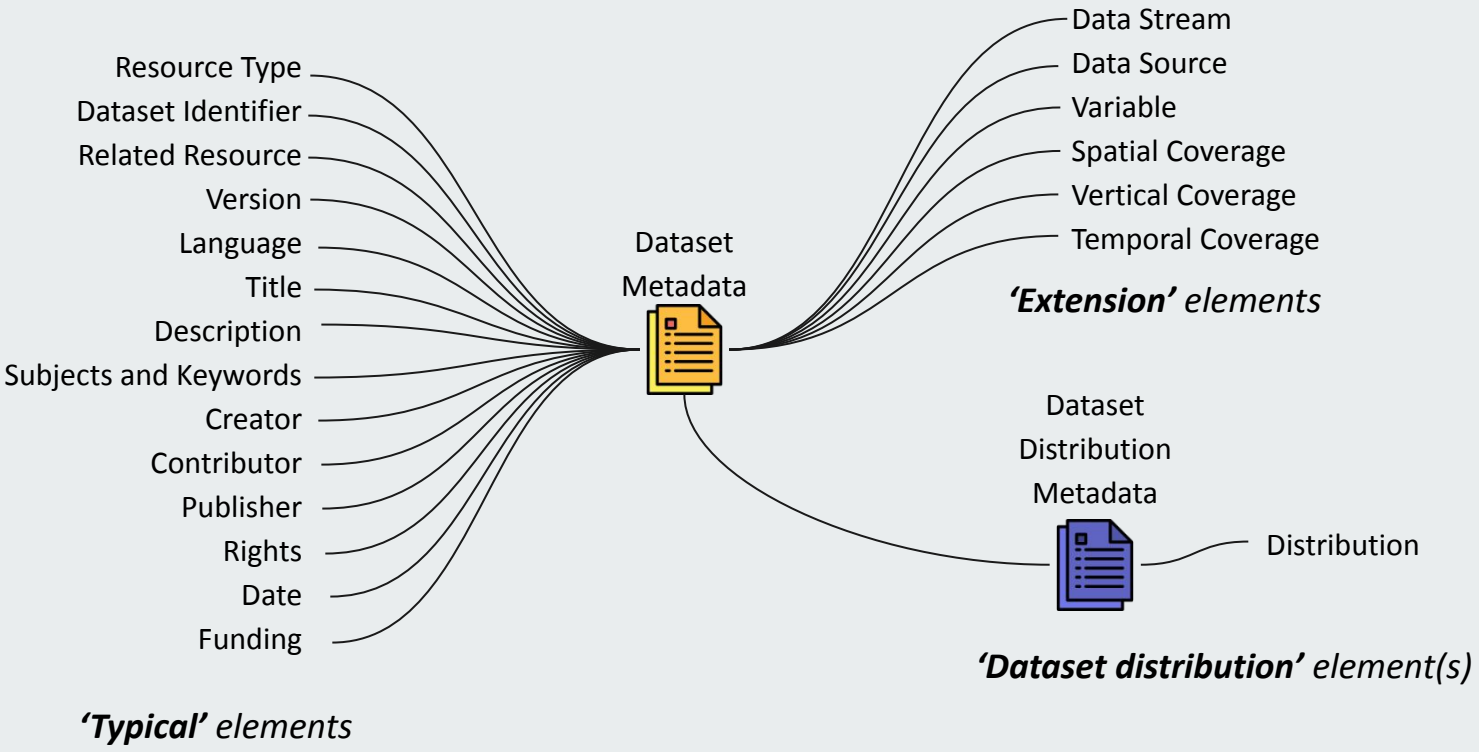
- Inspired by DataCite and DCAT scheme
- Scheme fused, improved, extended and ‘simplified’

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

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

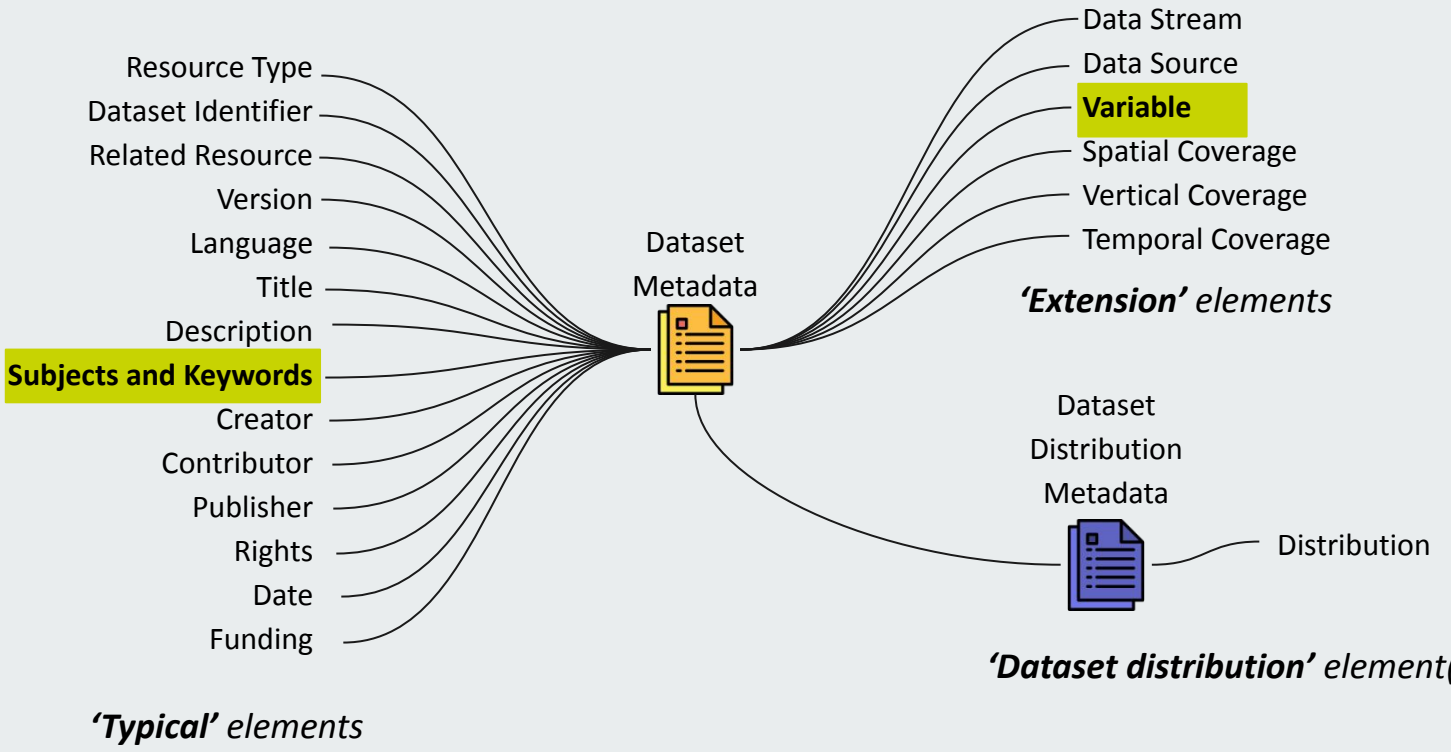
- GDMT contains a **‘back-end’ vocabulary** that enables machine-actionability, which contains:
  - ~**130** RDF properties
  - ~**1000** controlled terms

- **GDMT was started during M4M.5 & M4M.6**



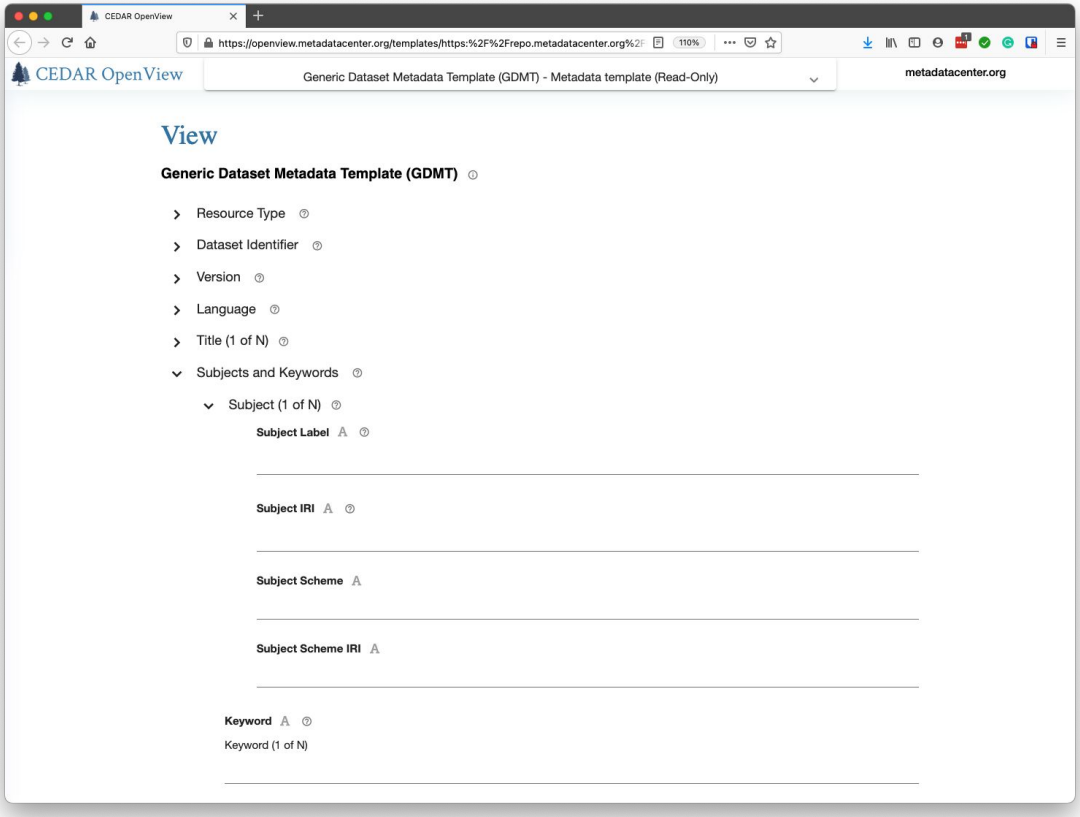
You can find definitions of elements and fields on CEDAR and GitHub.

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

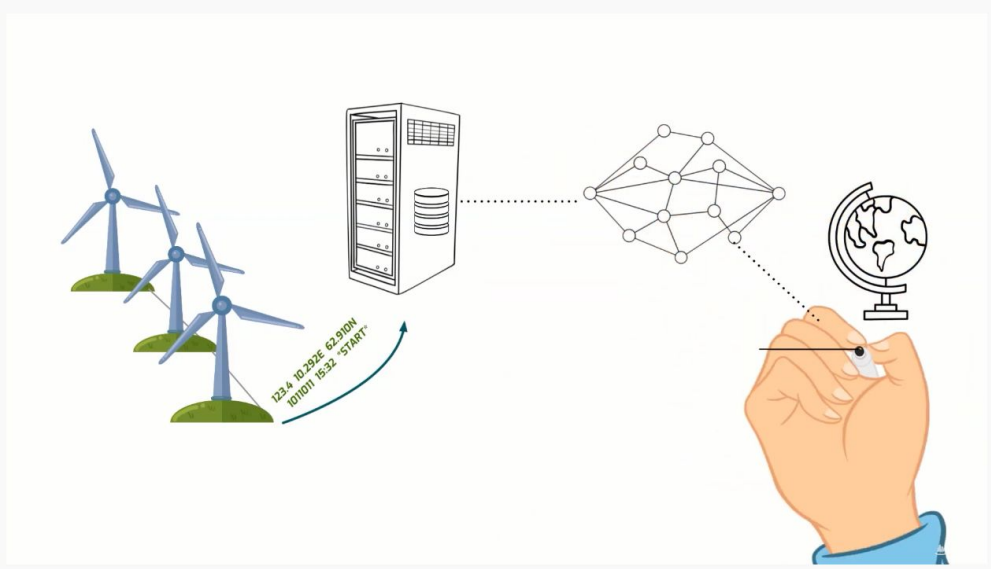


By creating domain specific controlled vocabularies and updating GDMT to use them, we turn this template to be domain specific

GDMT in CEDAR OpenView



See Wind Energy use case

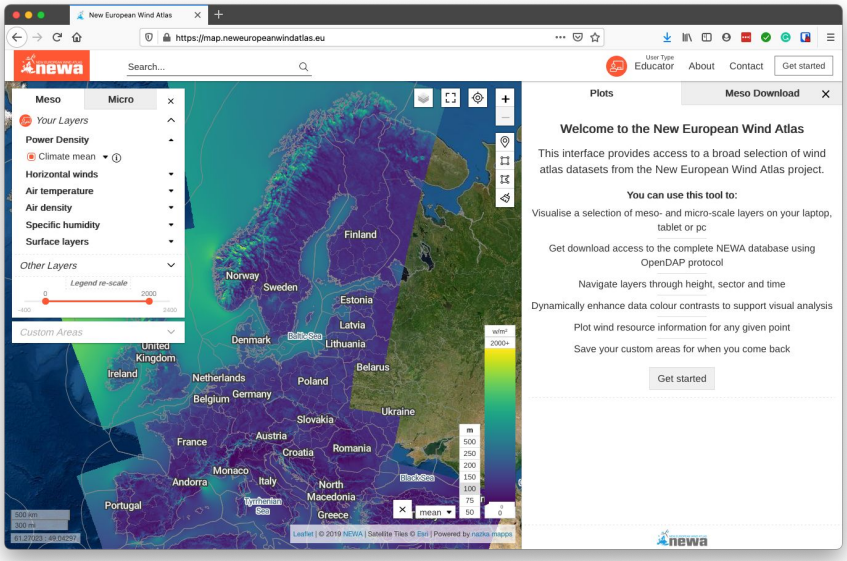


http://bit.ly/we-fair

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] micro scale data subsetting and aggregation



Thank you.

Source of graphical material




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