>PINK

Semantic data documentation for FAIR data

Jesper Friis, Francesca L. Bleken (SINTEF)

FAIRification workshop

St. Gallan, January 14, 2025

THE PINK PROJECT HAS RECEIVED FUNDING FROM THE EURO-PEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO. 101137809.

ASSOCIATED PARTNERS (I.E. (A) SWISS PARTNERS AND (B) UK PARTNERS) HAVE RECEIVED NATIONAL FUNDING FROM (A) THE SWISS STATE SECRETARIAT FOR EDUCATION, RESEARCH AND INNOVATION (SERI), AND (B) INNOVATE UK.

From raw data to knowledge exploitation

Raw data

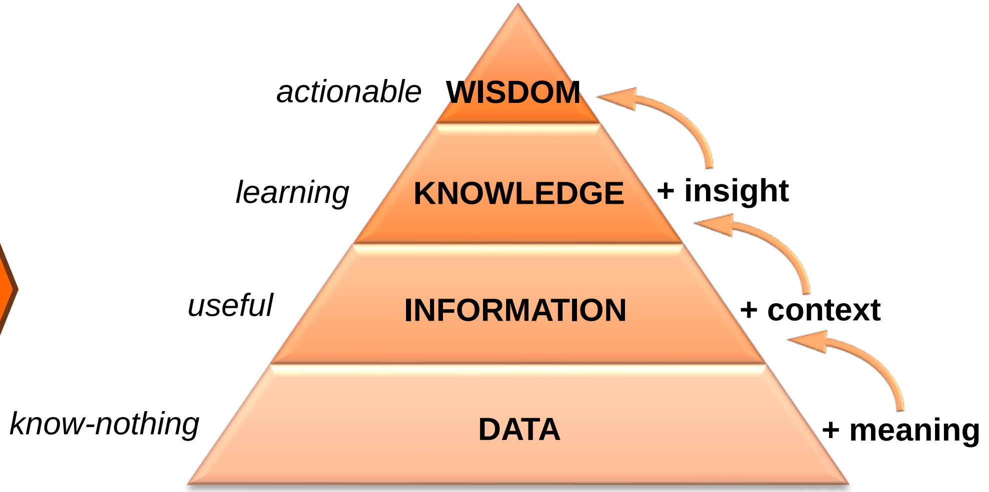
Model output

Characterisation results

Undocumented data source

Knowledge framework

Semantic enhancement



DIKW-pyramid

- Overview over available data
- FAIR data
- Traceability

Exploitable data

Semantic toolchain

Represent datasets the way you need them

Knowledge integration

Knowledge exploration

Levels of data documentation

Raw data

FAIR
Cataloguing
Where and how the data can be found and accessed

FAIR
Structural documentation
How the data is structured and represented numerically

DLite

Formalised data models

Data model
property 1
property 2

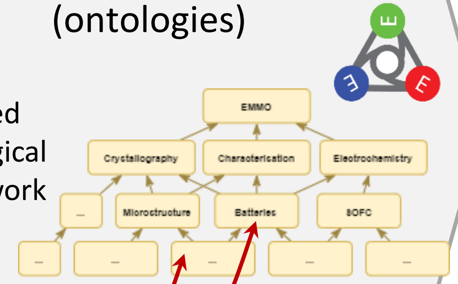
FAIR
Contextual documentation
How the data relate to other data and asserts (linked data)

JSON-LD

Enable contextual search and traceability

FAIR
Semantic documentation
What *is* the data. Fully described using a knowledge language (ontologies)

Shared ontological framework



Data model
property 1
property 2

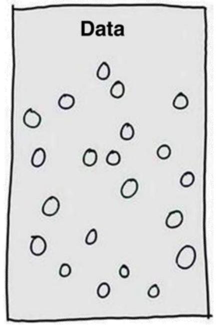
Semantic meaning via mappings

Exploitable data

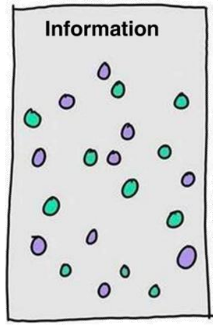
meaning

context

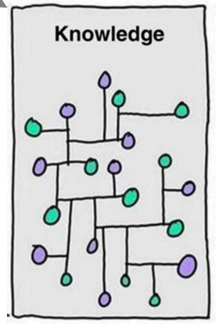
insight



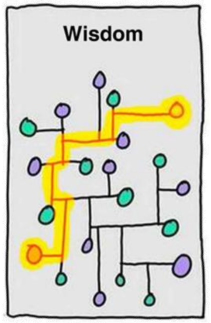
Signals
Contrast
Know-nothing



Useful
Meaningful
Organised
Structured



Contextual
Synthesised
Learning



Understanding
Integrated
Actionable

Cataloguing – facts about data

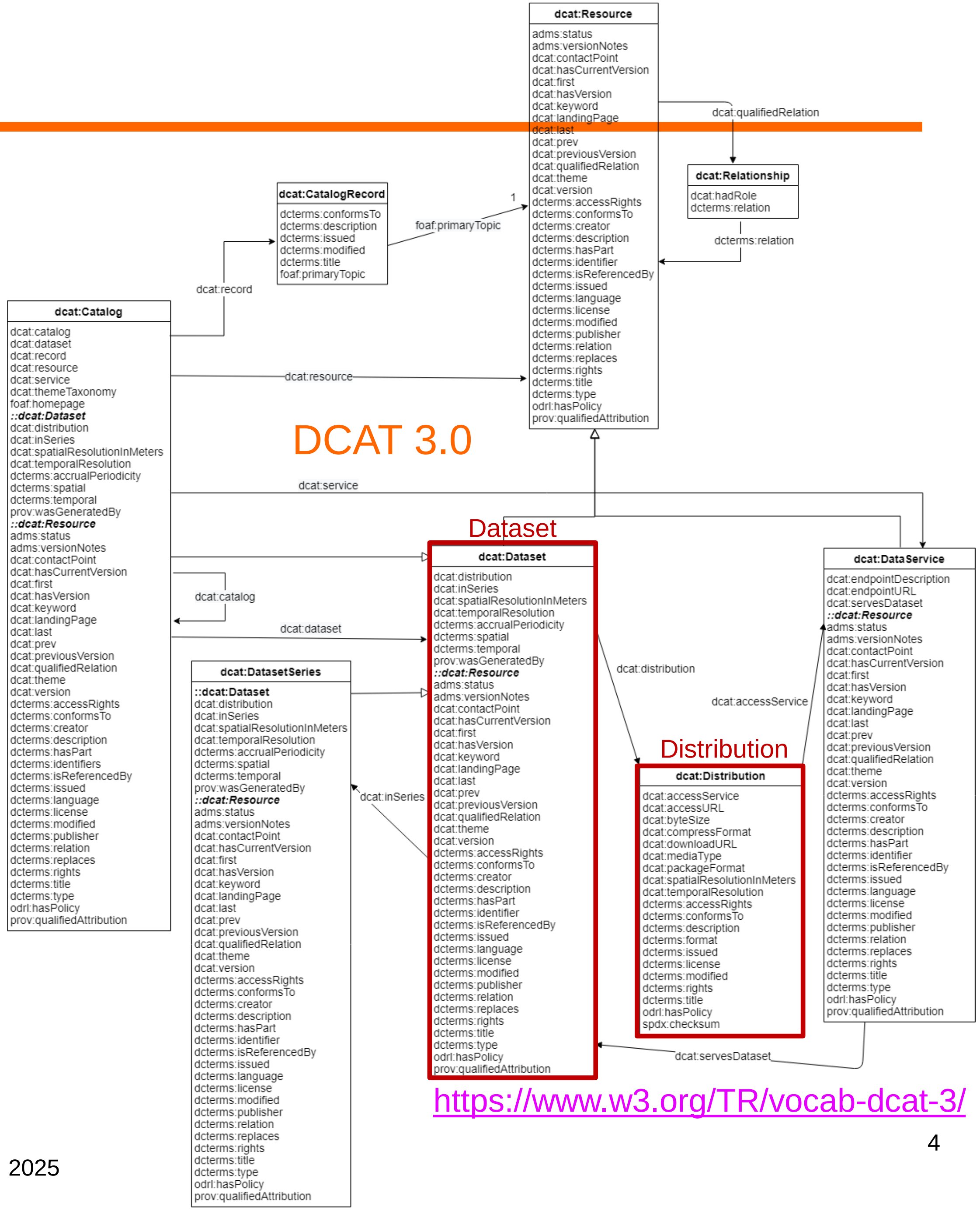
Provides: findability and accessibility

Also, makes data usable

Important to use **common standards** to make our data findable and accessible for external consumers

Datasets are central for cataloguing

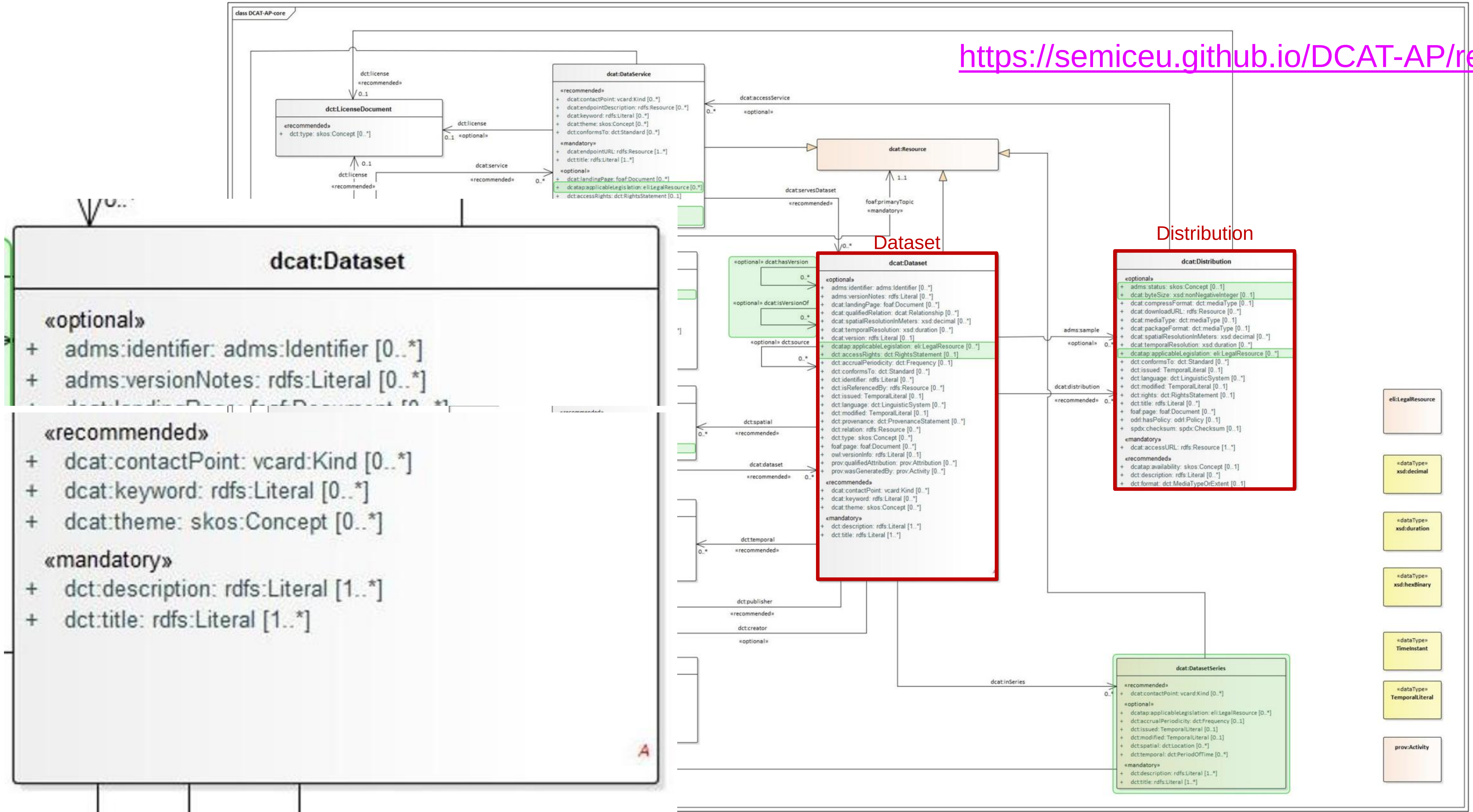
- DCAT
- DCAT-AP (European, adds additional requirements)



https://www.w3.org/TR/vocab-dcat-3/

DCAT Application Profile (DCAT-AP)

https://semiceu.github.io/DCAT-AP/releases/3.0.0/



Structural documentation – structure of data

Provides: interoperability (at low/numerical level)

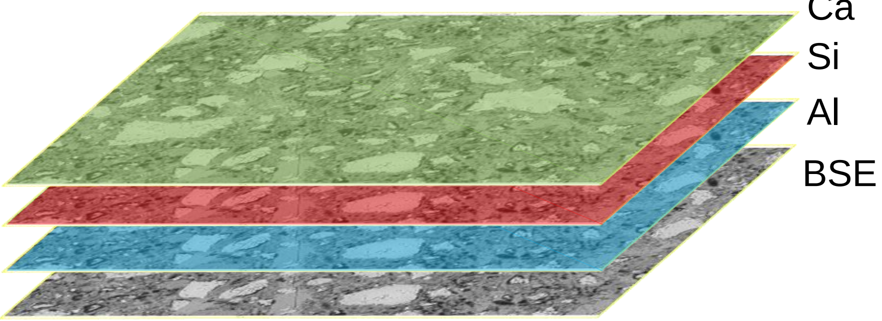
Datamodel	
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage
Description:	SEM image with elemental mappings. Represented as…

Dimensions	
Name	**Description**
nelements	Number of elements for elemental mapping.
height	Number of pixels along image height.
width	Number of pixels along image width.

Properties				
Name	**Type**	**Shape**	**Unit**	**Description**
data	float32	[nelements+1, height, width]		Image data
elements	string	[nelements]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref http://.../0.1/SEM			Ref. to data model for SEM metadata

Formalised data models used to describe the structure of datasets

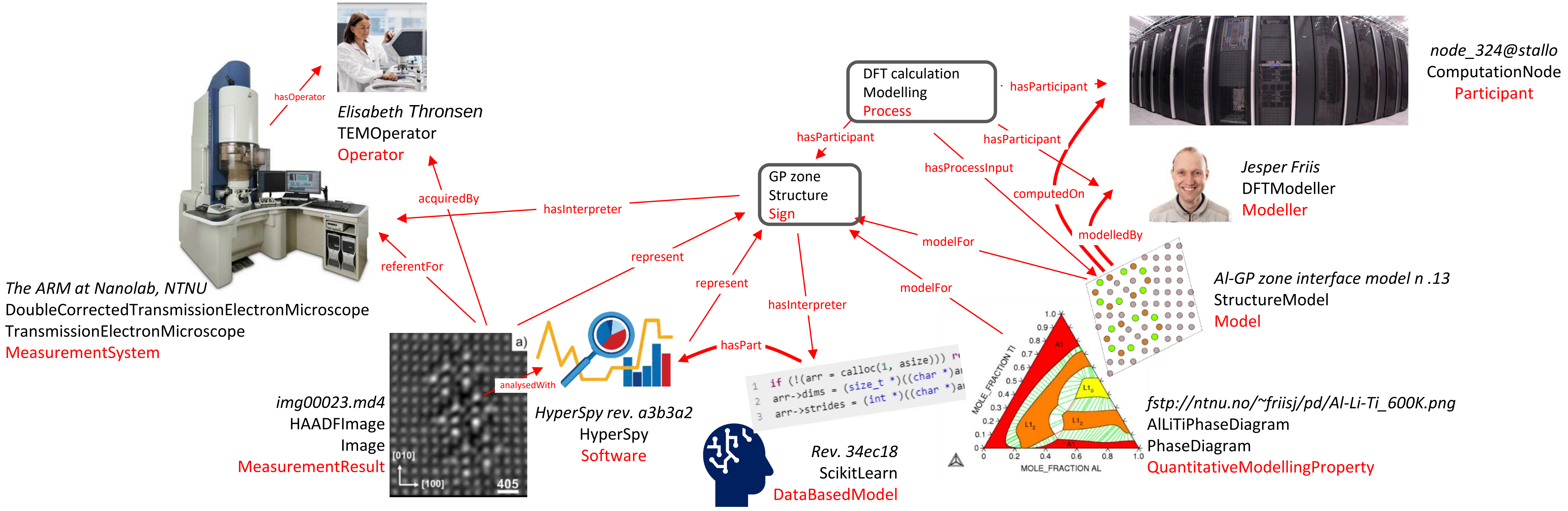
The described dataset
SEM image with elemental mapping



Contextual documentation → knowledge base

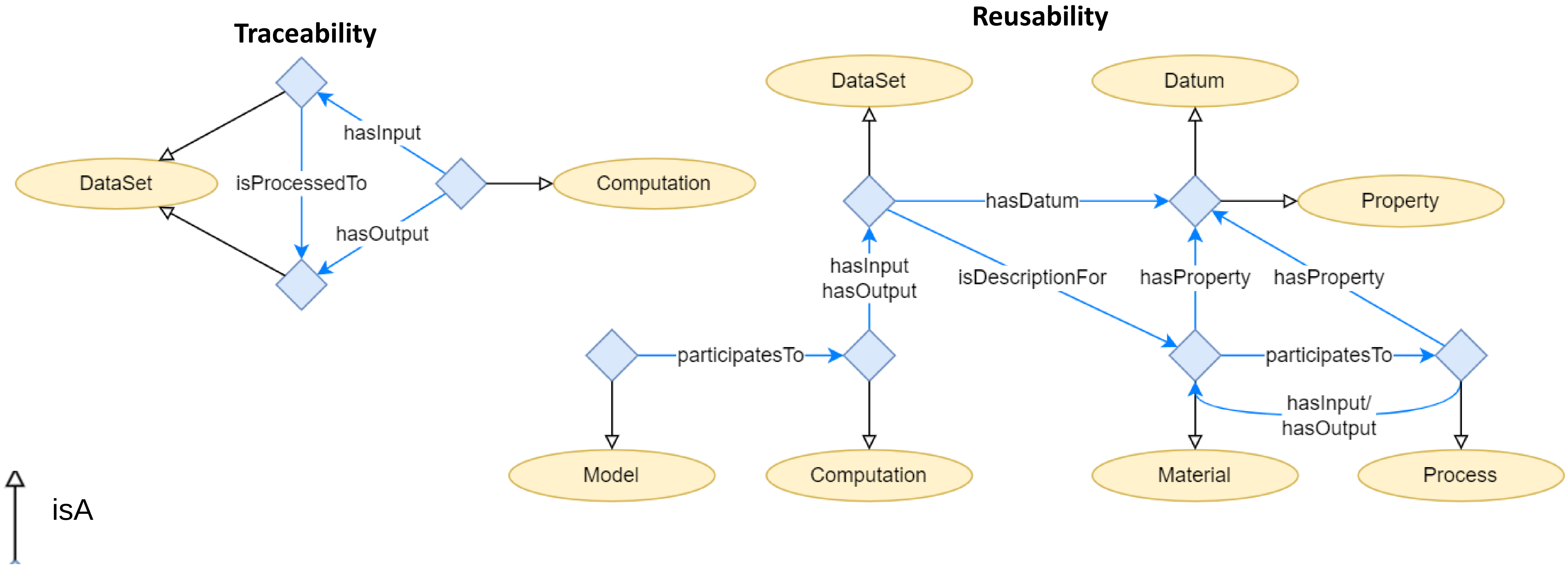
Provides: context (reusability)

Express relations between different documented resources



Contextual documentation → knowledge base

Palette of basic relations between different types of resources (datasets, materials, properties, models…)



Semantic documentation → autonomous systems

Provides: actionable data

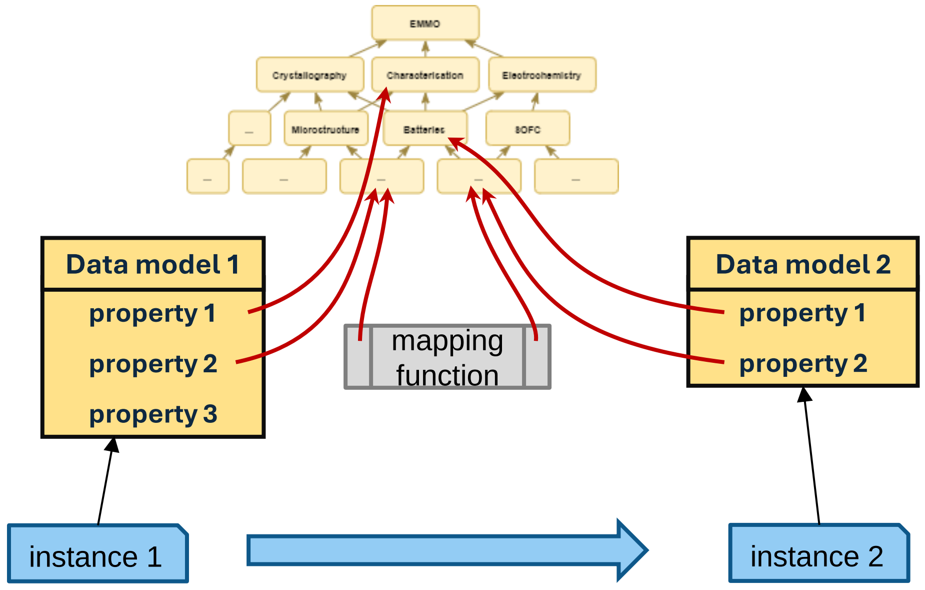
What **is** the documented resource

Enhances

- findability (semantic search)
- interoperability (semantic interoperability, cross-domain interoperability)
- reusability

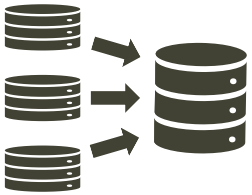
Represent datasets the way you need them

Mapping datamodels to ontological concepts

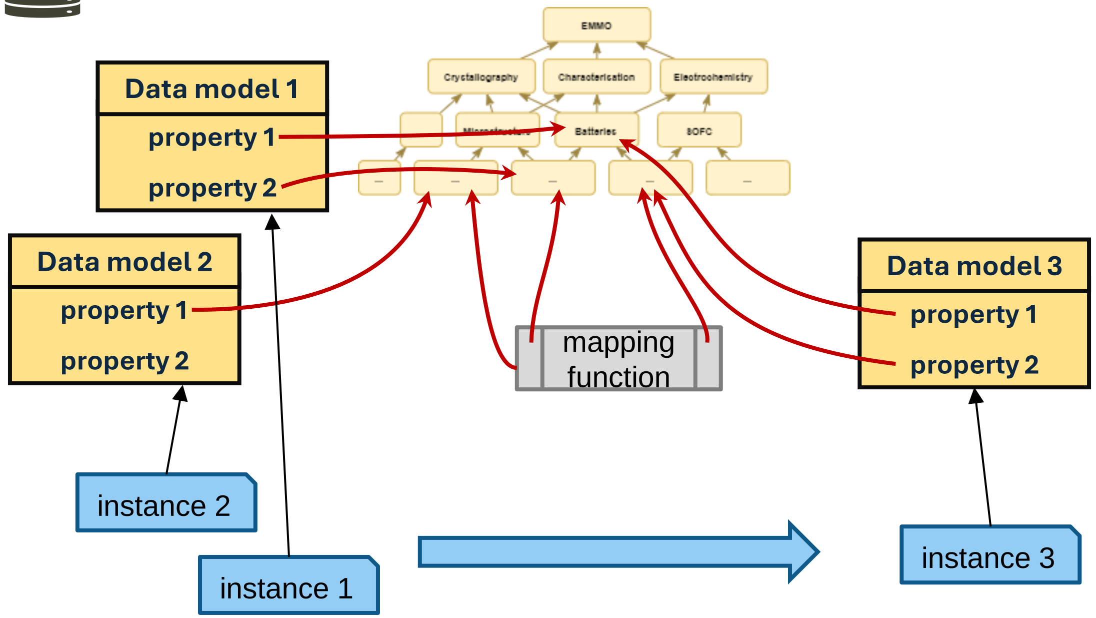


Convert between representations

Semantic documentation → autonomous systems

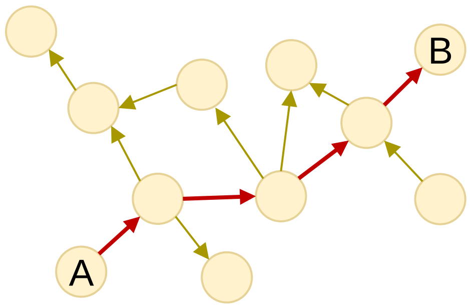


Knowledge integration



System that automatically can combine datasets

Knowledge exploration



Use the knowledge graph to explore how to get from A to B

Example: The OntoFlow workflow builder in OpenModel

Tooling

Tripper

rdflib (in-memory)

GraphDB

Apache Jena Fuseki

Stardog

…

Consistent Python interface to various triplestore backends

tripper.datadoc

Tripper subpackage for data documentation

Interfaces

- Python API
- Command-line tool
- Web interface (SOFT Studio)

Documentation input

- Python dict
- YAML file
- CSV file
- Excel
- Web form

DLite

A lightweight data-centric framework for semantic interoperability

Formalised datamodels

For structural documentation and interoperability

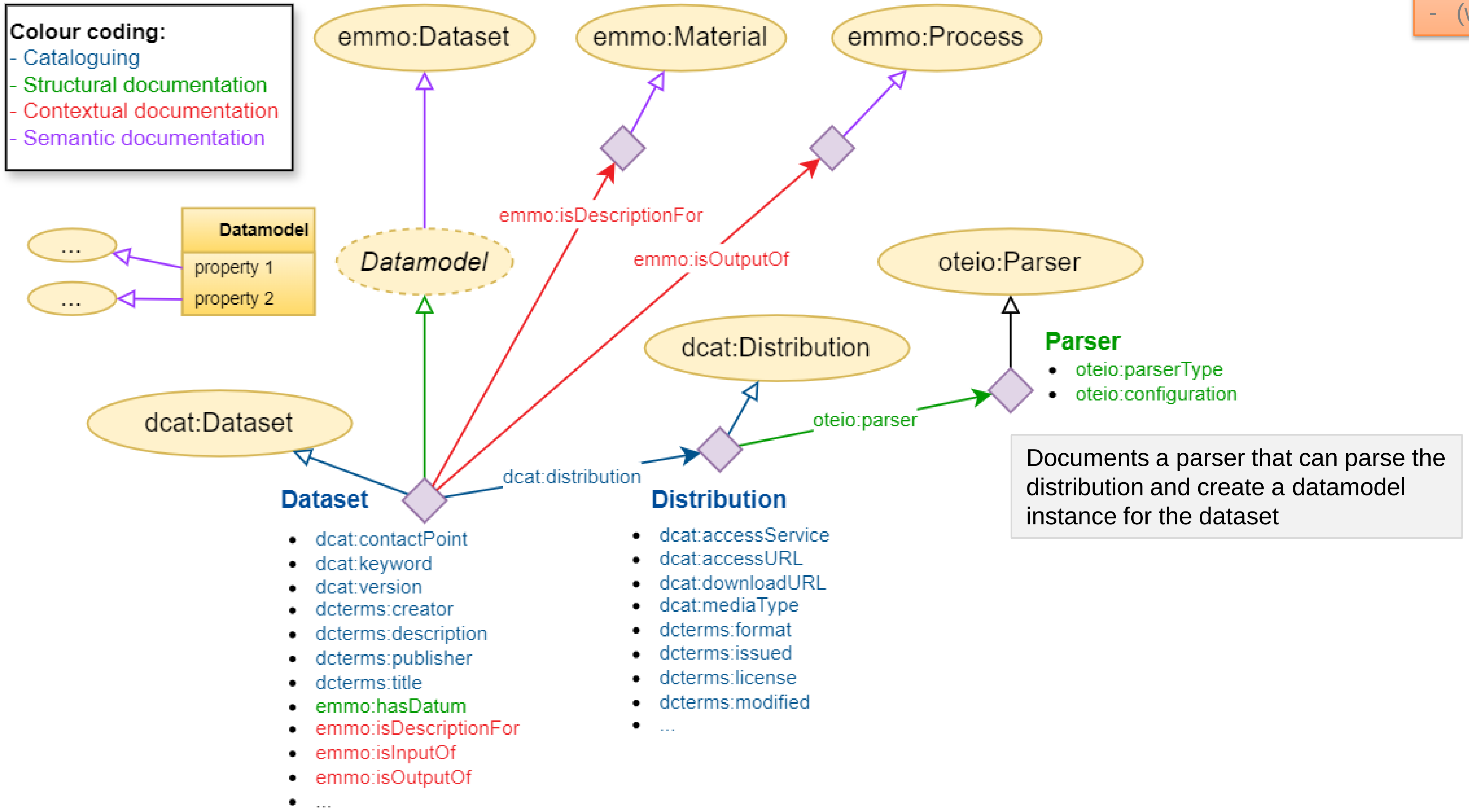
Datamodel				
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage			
Description:	SEM image with elemental mappings. Represented as…			
Dimensions				
Name	**Description**			
nelements	Number of elements for elemental mapping.			
height	Number of pixels along image height.			
width	Number of pixels along image width.			
Properties				
Name	**Type**	**Shape**	**Unit**	**Description**
data	float32	[nelements+1, height, width]		Image data
elements	string	[nelements]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref http://.../0.1/SEM			Ref. to data model for SEM metadata

Roles

- **Data provider**
  - Document and make **existing** datasets accessible
  - Ex: experimentalist/characterisation instrument, modeller, …
- **Data user**
  - Access and use **existing** (documented) data
- **Data consumer**
  - Document what type of data is needed (the data may **not** exist (yet))
  - Ex: model (input)
- **Data producer**
  - Document what type of data is produces (the data may **not** exist (yet))
  - Ex: model (output)

Data provider – documenting a dataset

Can be expressed in:
- Python
- YAML
- CSV
- (web form)



Data user

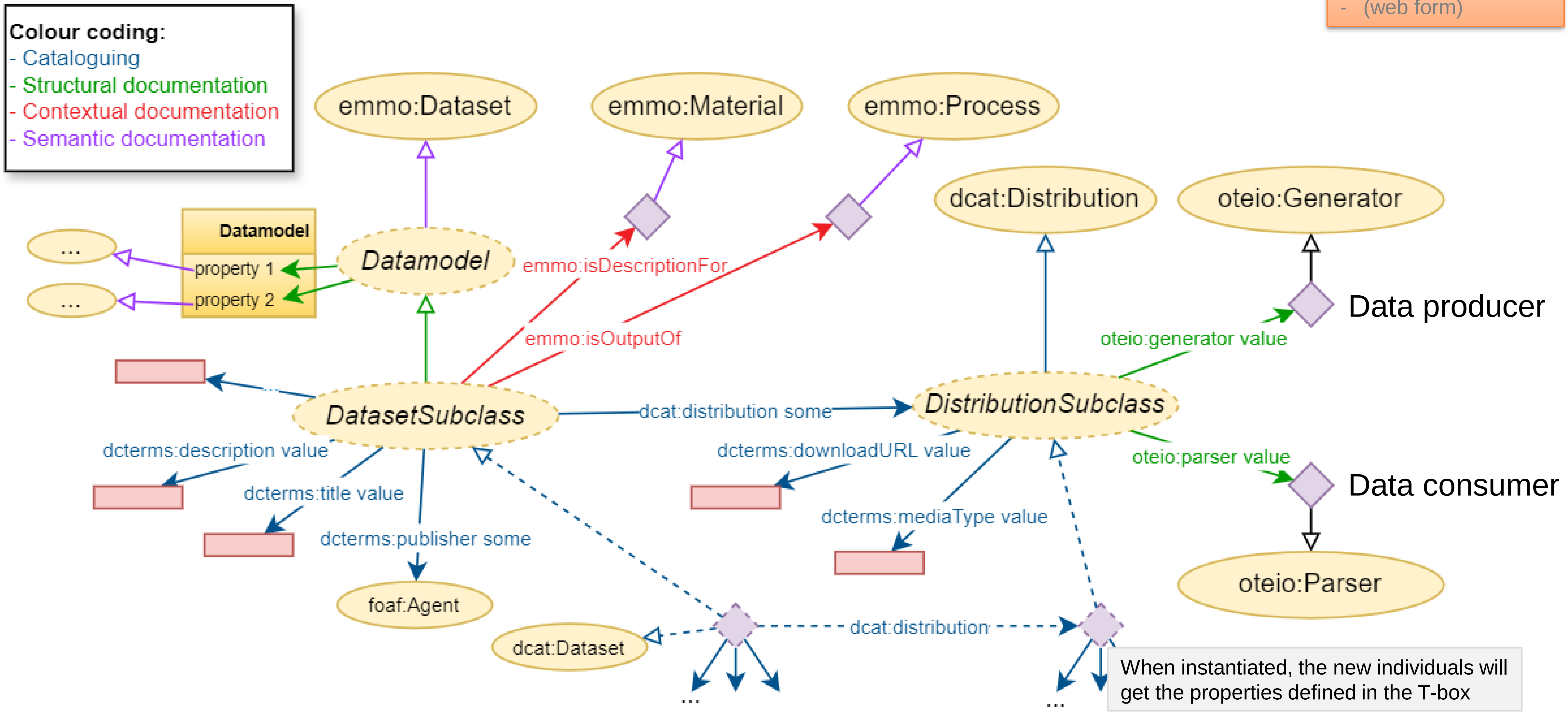
Can be expressed in:
- Python
- datadoc tool
- (web interface)

- Find data
  - search criteria → IRI of matching resources
- Access data
  - provide the IRI
    → raw data
  - provide IRI(s) + description of how the data should be represented
    → data represented the way you want it

- Do not have to add any documentation to the knowledge base

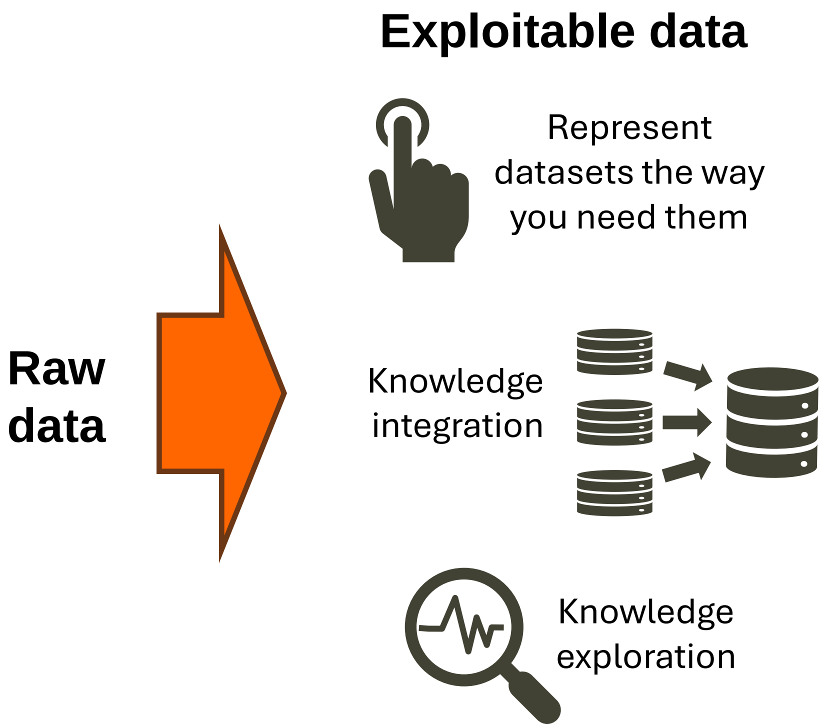
Data consumer & data producer (T-box)

Planned input:
- Python
- YAML
- CSV
- (web form)



Summary

- Data documentation framework
  - enable knowledge exploitation
  - support all levels of data documentation
    - cataloguing
    - structural documentation
    - contextual documentation
    - semantic documentation
  - supported user roles
    - data provider (document existing data)
    - data user (access and use existing documented data)
    - data consumer (ex: model input…)
    - data producer (ex: model output…)
  - Software packages: Tripper, DLite, SOFT Studio (to be implemented in PINK)



>THANK YOU!

Acknowledgement:

This work has received funding from the *European Union's Horizon 2020 research and innovation programme* and Norwegian Research Council via the following projects:

- OntoTrans (2020-2024) Grant Agreement n. 862136
- SFI PhysMet (2020-2028) Research Council of Norway n. 309584
- OpenModel (2021-2025) Grant Agreement n. 953167
- NanoMECommons (2021-2025) Grant Agreement n. 952869
- MatCHMaker (2022-2026) Grant Agreement n. 101091687
- PINK (2024-2028) Grant Agreement n. 101137809

WWW.PINK-PROJECT.EU

THE PINK PROJECT HAS RECEIVED FUNDING FROM THE EURO-PEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO. 101137809.

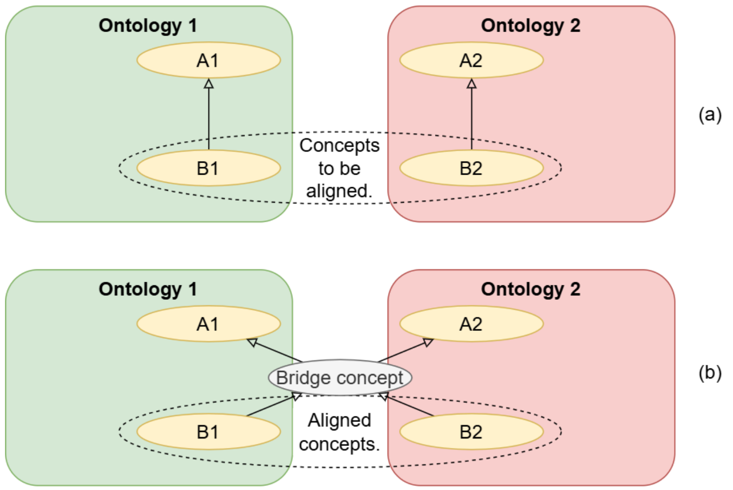
ASSOCIATED PARTNERS (I.E. (A) SWISS PARTNERS AND (B) UK PARTNERS) HAVE RECEIVED NATIONAL FUNDING FROM (A) THE SWISS STATE SECRETARIAT FOR EDUCATION, RESEARCH AND INNOVATION (SERI), AND (B) INNOVATE UK.

Herfra er det bare gamle slides

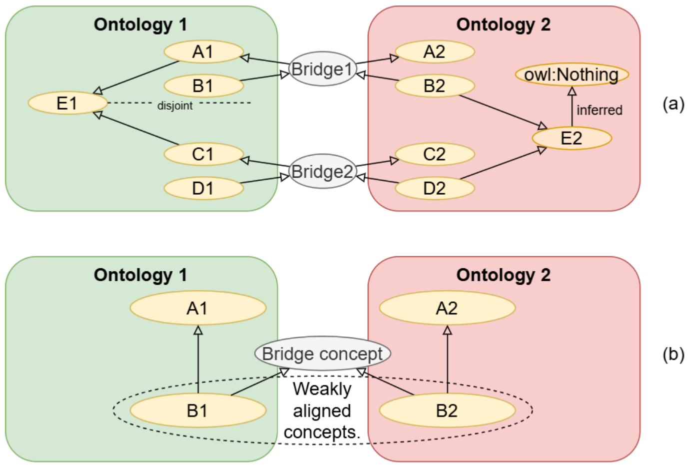
Ontological alignment

Important when combining different ontologies
(often needed for inter-disciplinary problems)

Bridge concepts

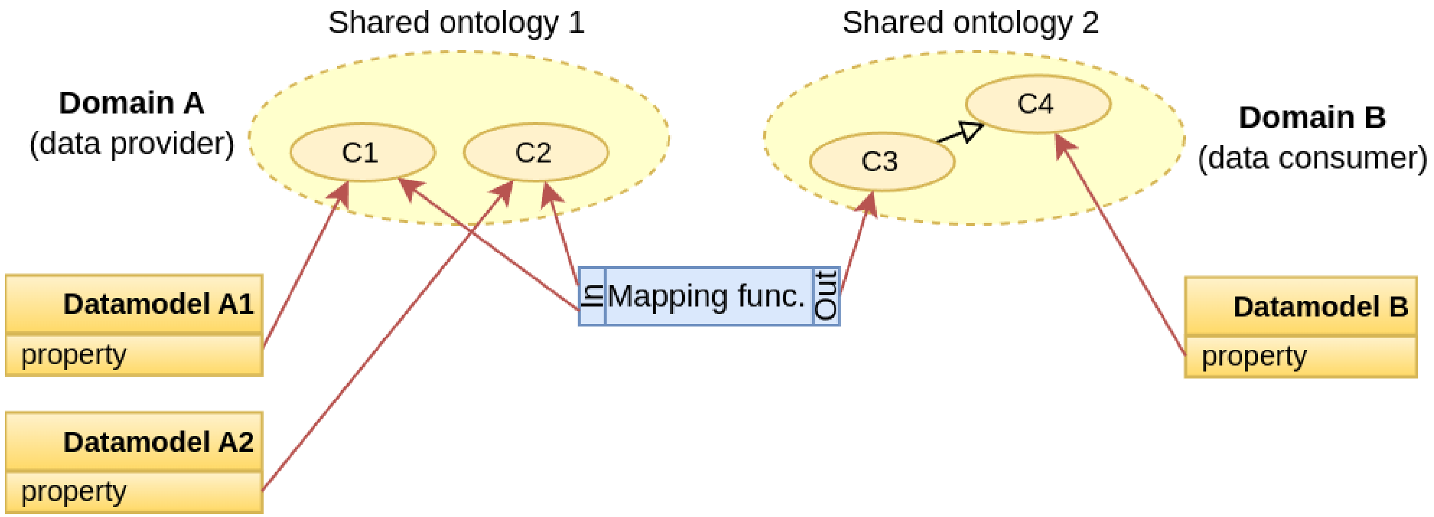


Weak bridge concepts



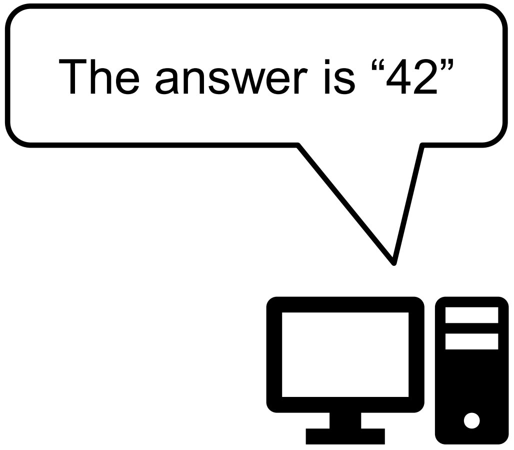
Ontological alignment

Mapping functions can also be used for ontological alignment



Contextual documentation

Provides: context (reusability)



What is the question?

What is documented

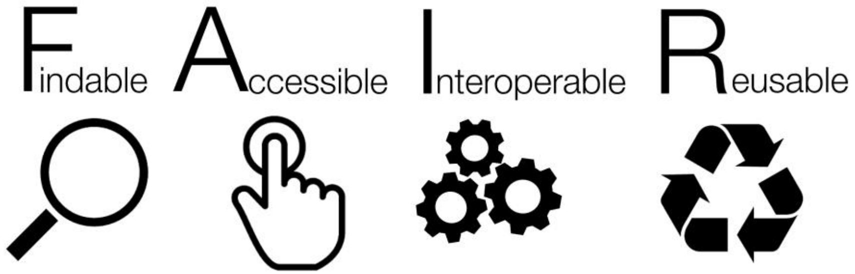
Roles

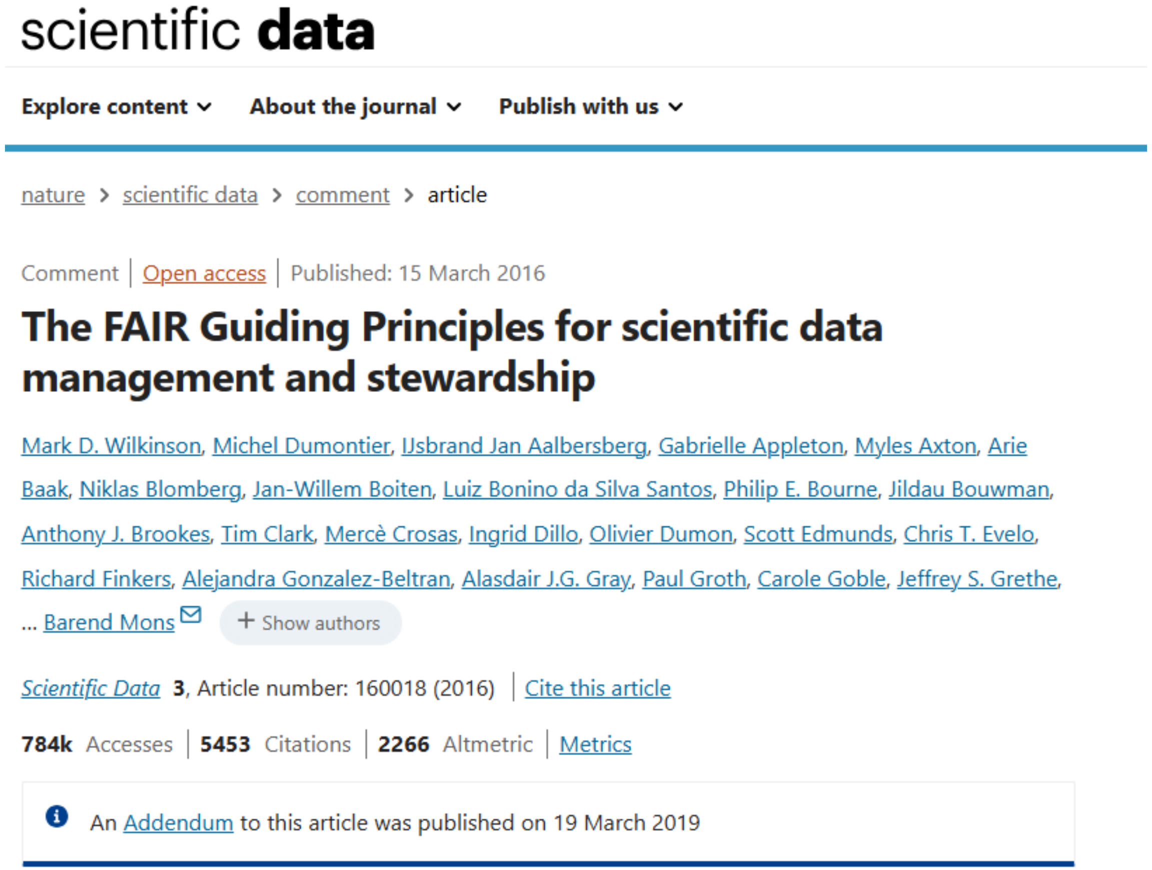
- **Data provider**
  - Document and make **existing** datasets accessible
  - Ex: experimentalist/characterisation instrument, modeller, …
- **Data user**
  - Access and use **existing** (documented) data
    - get data as is
    - get standardised data (allows conversions)
    - get semantic data (
- **Data consumer**
  - Document what type of data is needed (the data may **not** exist (yet))
  - Ex: model (input)
- **Data producer**
  - Document what type of data is produces (the data may **not** exist (yet))

Overview

- Introduction to FAIR data documentation
- Ontologies
- Implementation

What is FAIR research data?





scientific **data**

Explore content ⌄ **About the journal** ⌄ **Publish with us** ⌄

nature › scientific data › comment › article

Comment | Open access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, … Barend Mons + Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

784k Accesses | **5453** Citations | **2266** Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Widely endorsed

- 2016: **G20 leaders** (at the G20 Hangzhou summit)
- 2016: **EU**: H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 Version 3.0
- 2017: **Regjeringen**: Nasjonal strategi for tilgjengeliggjøring og deling av forskningsdata
- 2018: Association of European Research Libraries
- 2021: UNESCO Recommendation on Open Science
- …

Important aspects

- Persistent identifiers
- Adequate metadata and documentation
- Clear licenses
- Standards and formats

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Are the FAIR principles sufficient?

General guiding principles

… but the implementation depends on the needs

Ex, data exchange within a single domain or cross-domain

- Quantum chemistry
- Fluid dynamics
- Nano safety
- Life cycle analysis
- Circular economy
- …

SINTEF

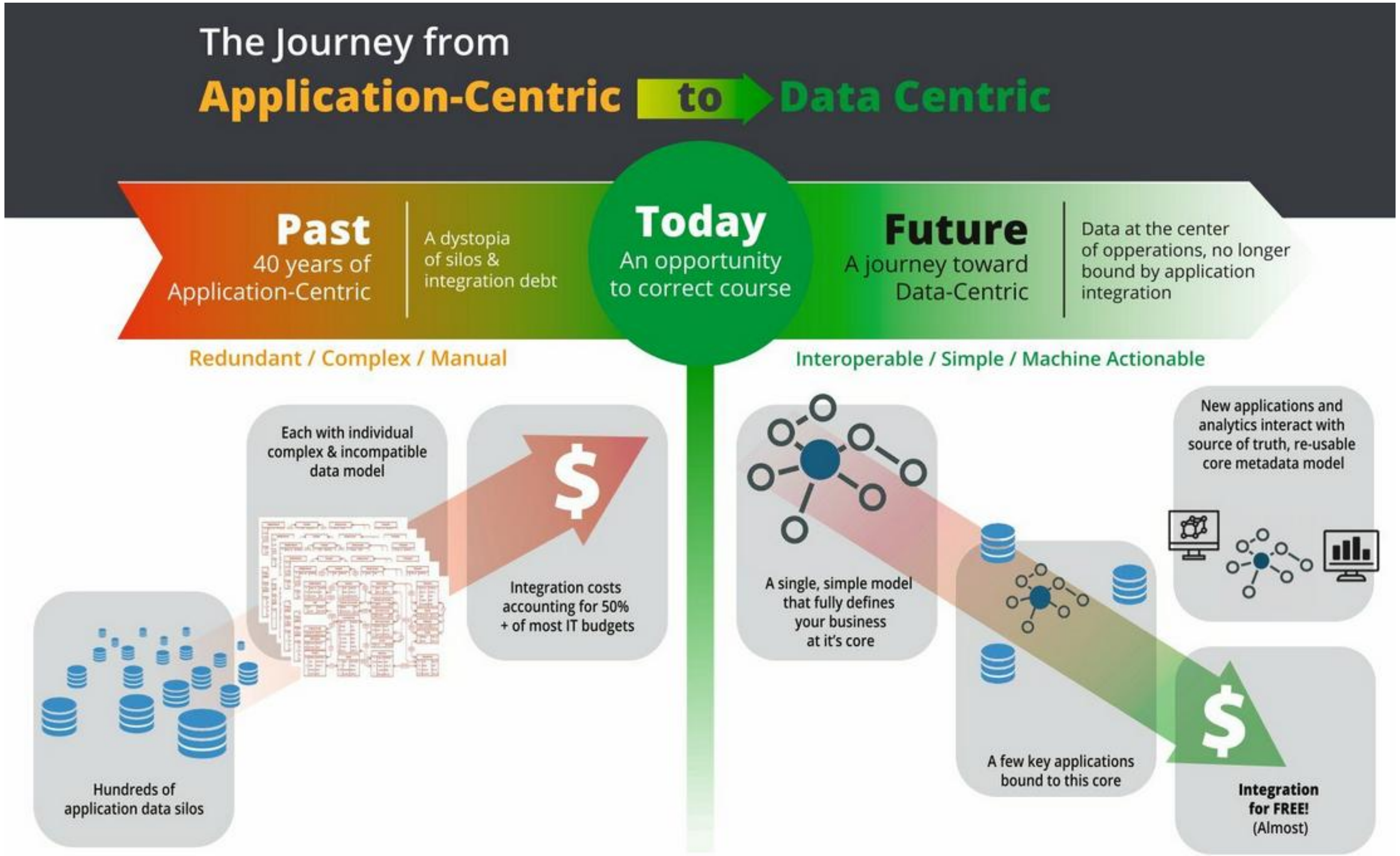
Trends:

Data Floods and Meaning Crisis

Images generated by Microsoft Copilot Designer. G. Goldbeck

SINTEF

Data centricity – structure data around science rather than IT systems

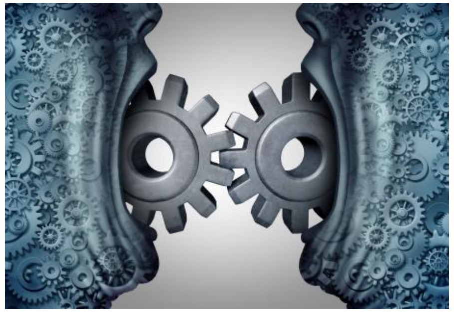


From Semantic arts;
https://www.semanticarts.com/

SINTEF

“Data-centric” really *means* “Concept-centric”

- *Structure data around the **science*** (Ben Gardner, AstraZeneca)
- *Start with **meaning*** (Dave McComb, Semantic Arts*)
- *Start with **conceptualisation*** (Emanuele Ghedini, Univ Bologna, EMMO)
    - Express general scientific and domain knowledge
    - Utilise agreed standards

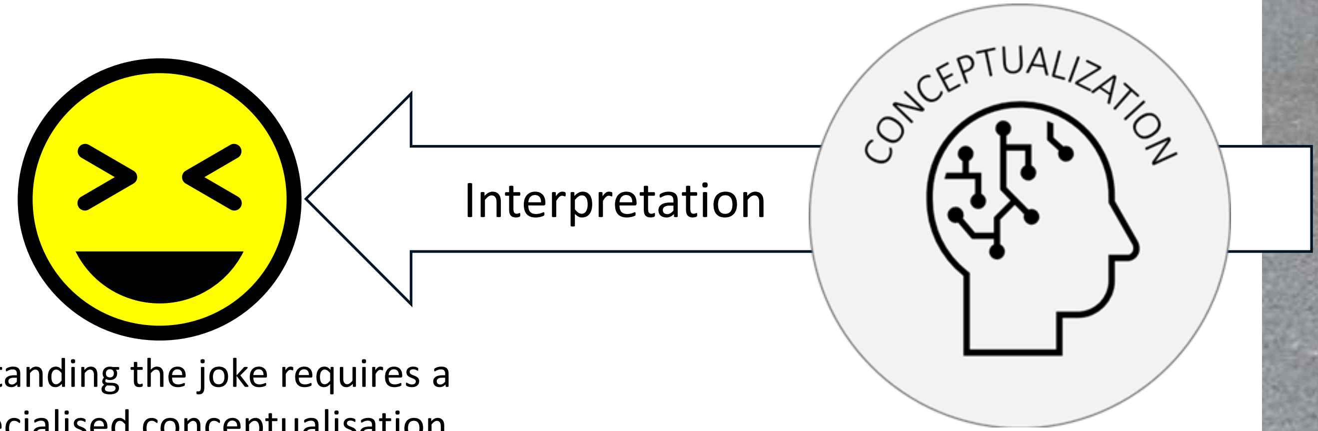


[* https://www.brcommunity.com/articles.php?id=b972](https://www.brcommunity.com/articles.php?id=b972)

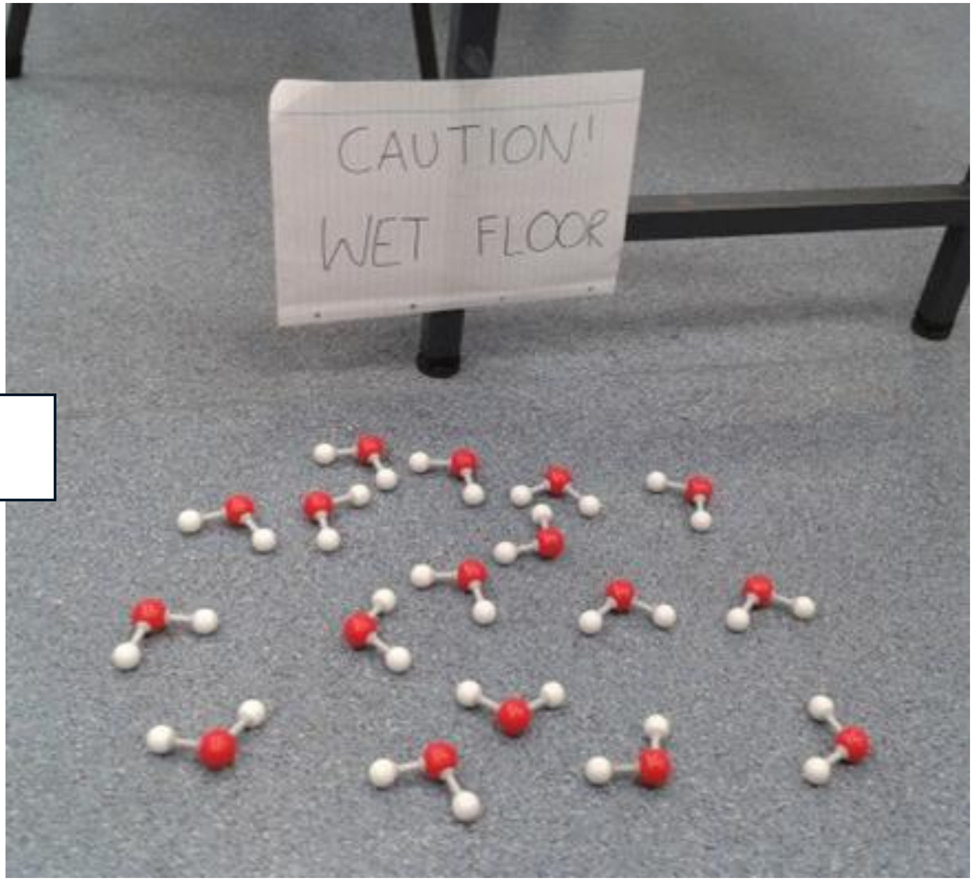
SINTEF

Conceptualisation

The interpretation of the world by someone



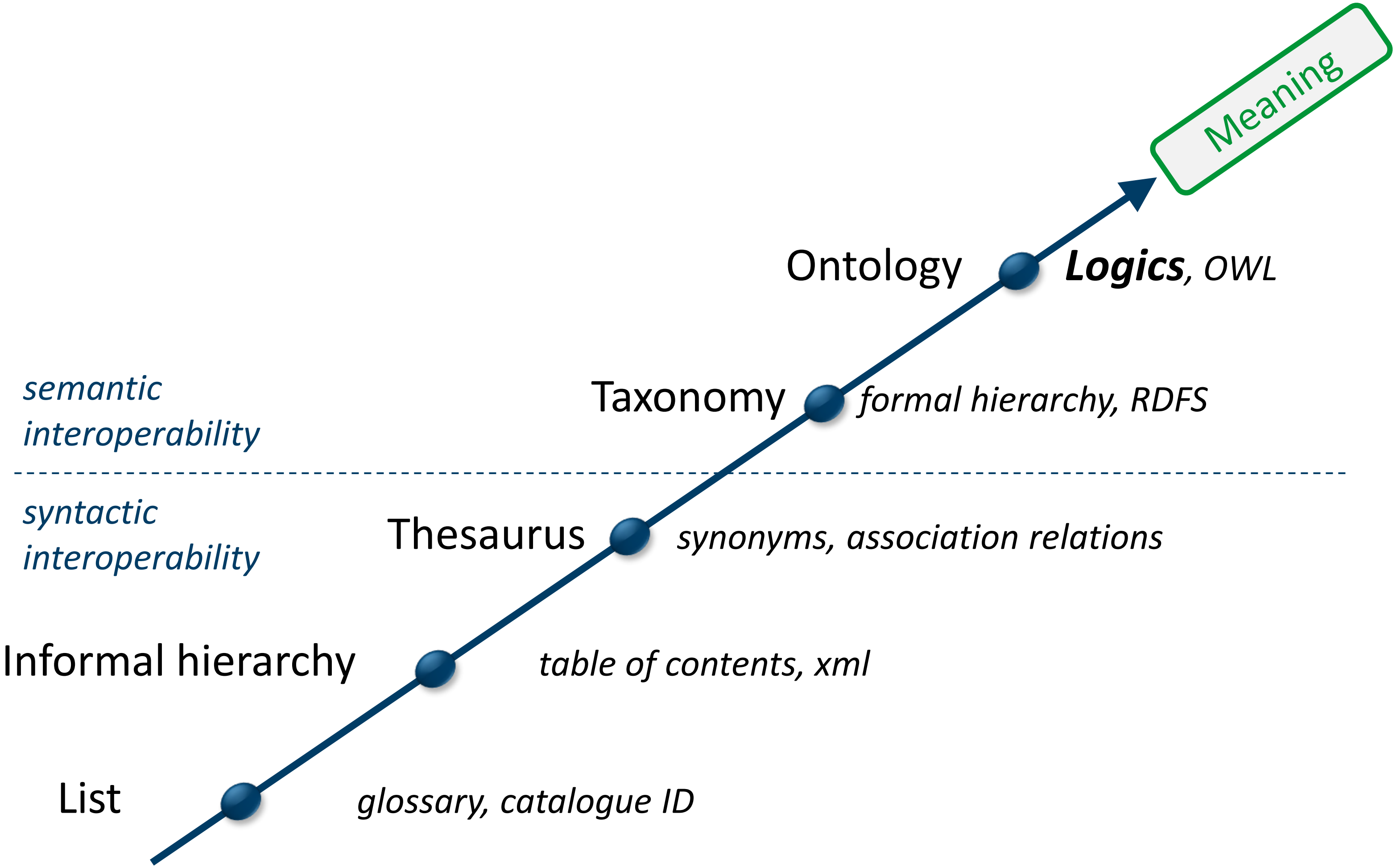
Understanding the joke requires a very specialised conceptualisation.



This image is just a bunch of pixels, unless interpreted by someone able to recognise that the balls stands for water molecules.

Data gets a **meaning** only when the data producer and consumer share the same conceptualisation.

Knowledge Organisation with Semantics



Machine can **interpret** information and **reason**.

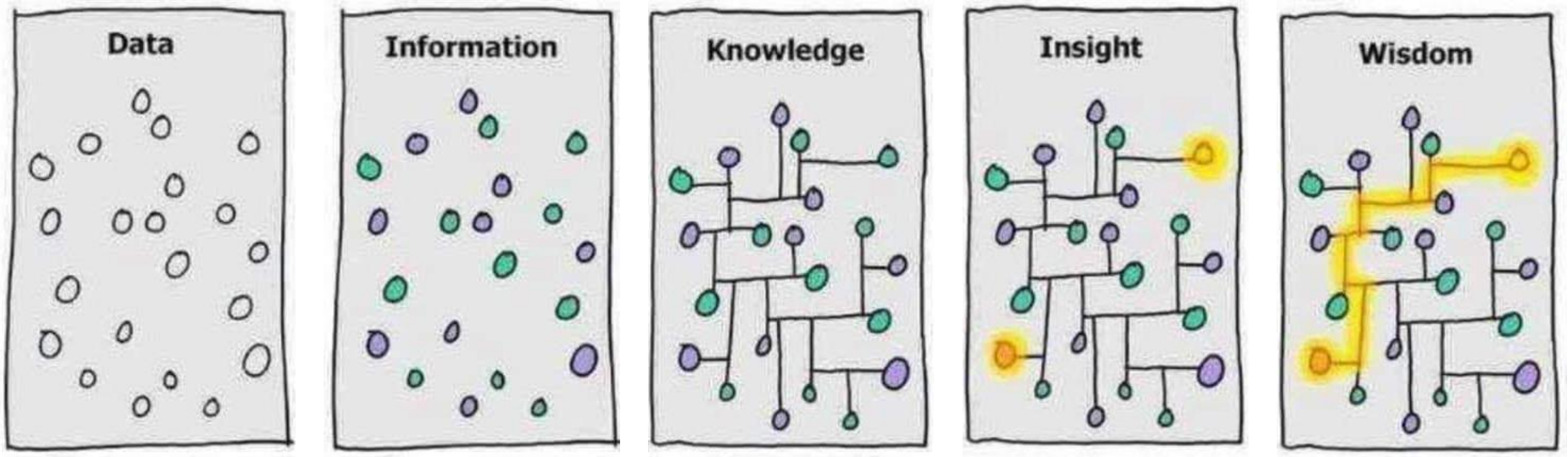
Machine can process information due to **compatible syntax**

Adapted from: Leo Obrst "The Ontology Spectrum". Book section in of Roberto Poli, Michael Healy, Achilles Kameas "Theory and Applications of Ontology: Computer Applications". Springer Netherlands, 17 Sep 2010.

From data to wisdom

Labels on data (*Information*) and connection between data (*Knowledge*) result from **Conceptualisation**

Insight and *Wisdom* follow with **semantic technologies**



symbols | who, what, where… | how | why | evaluated insight

SINTEF

Levels of data documentation

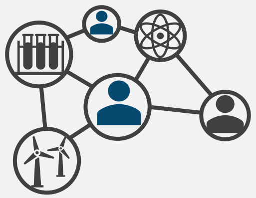
… and what aspects of FAIR data they address

FAIR

Cataloguing
Where and how datasets can be found and accessed

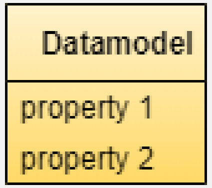
FAIR

Contextual documentation
How datasets relate to other data and things (linked data)



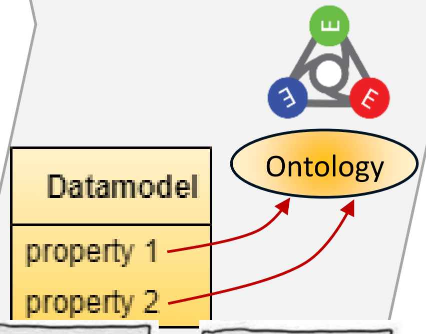
FAIR

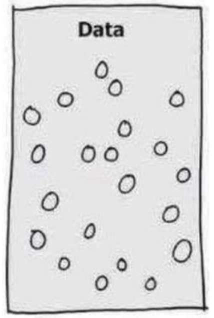
Structural documentation
How datasets are structured and represented numerically



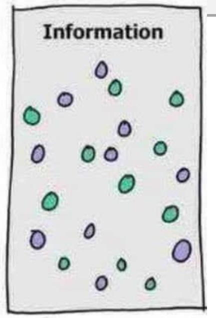
FAIR

Semantic documentation
What is the meaning of the data

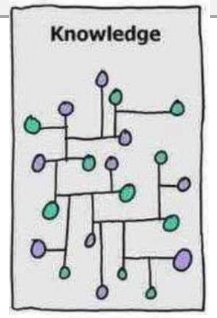




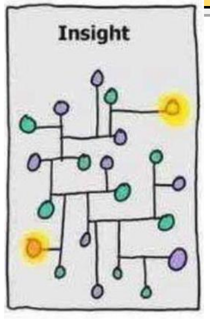
symbols



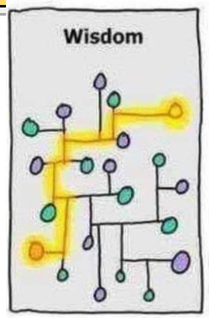
who, what, where…



how



why



evaluated insight

Intermezzo

Ontologies

What is an ontology

Ontology Entities

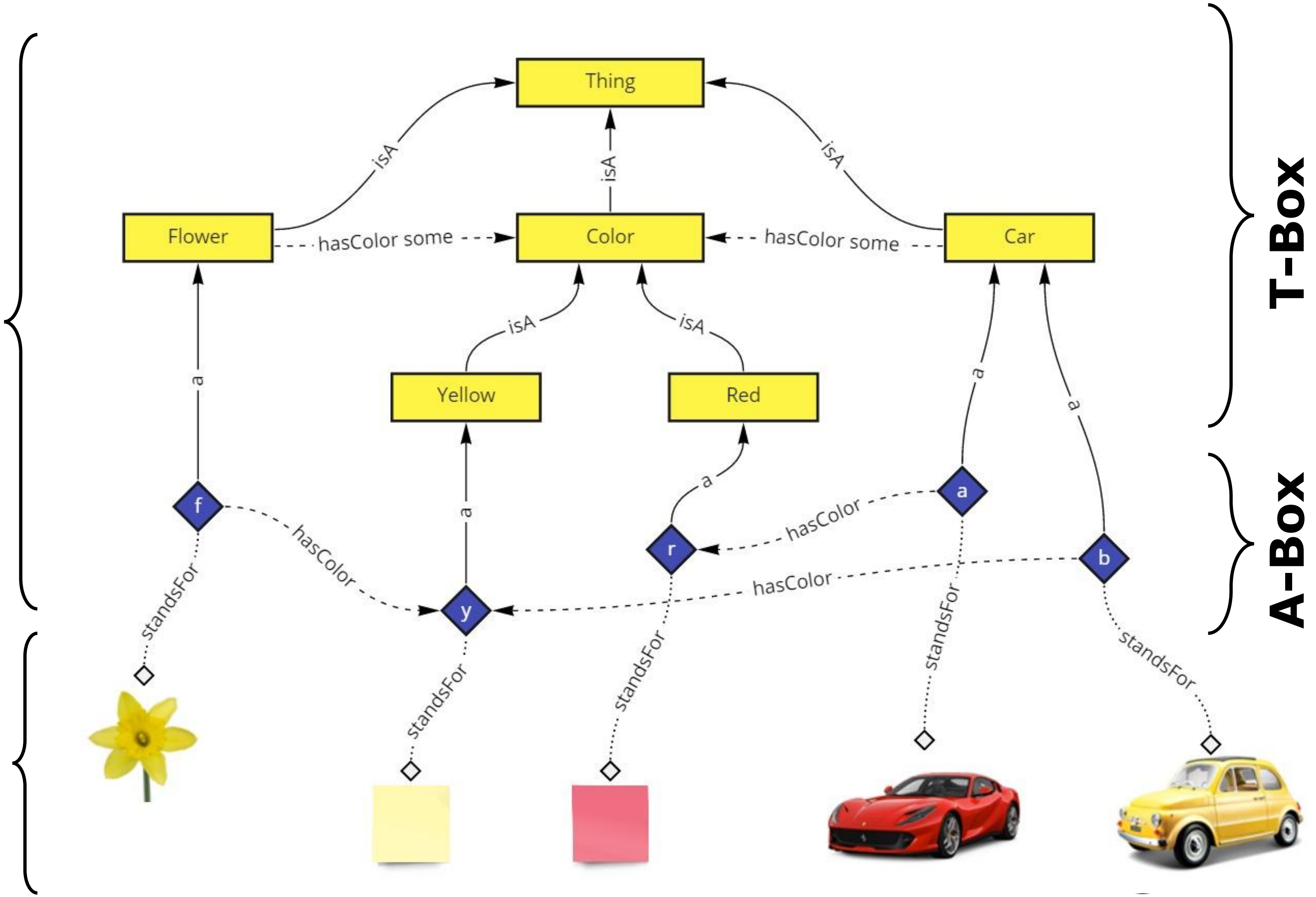
class relation individual

Example of axioms:

```
ClassAssertion( :a :Car)
ClassAssertion( :r :Color)
ObjectPropertyAssertion( :hasColor :a :r )
```

Real-world objects

No specific ontological commitment about the meaning of 'real' and 'object' in OWL 2. Relying on common sense.



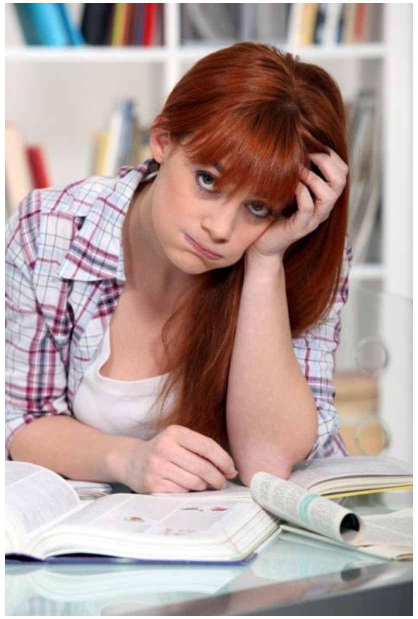
Benefits of Ontologies

- **Enhanced knowledge sharing and reusability** facilitate reuse across systems and applications
- Enhanced accuracy and **consistency** with formal representation of concepts and relationships
- Enhanced **interoperability** with shared vocabulary and conceptual framework
- Improved search and **discovery** in data source queries with structured representation of knowledge
- Facilitate automated **reasoning** to draw logical inferences based on relationships defined within the ontology

SINTEF

Materials Terminologies, Ontologies and Metadata Status

- Lack of agreed conceptualisation and terminologies (no equivalent to IUPAC Goldbook for Materials: even basic terms such as material vs chemical substance are taxonomically not clear.
- Poor coverage of materials domains in existing ontologies
- Lack of curation infrastructure
- Inconsistent use of TLO concepts when aligning key concepts such as materials property (e.g. quality, disposition etc).
- Low FAIR score of many semantic assets.



Review and Alignment of Domain-Level Ontologies for Materials Science
Anne De Baas, Pierluigi Del Nostro, Jesper Friis, Emanuele Ghedini, Gerhard Goldbeck, Ilaria Maria Paponetti, Andrea Pozzi, Lan Yang, Arkopaul Sarkar, Francesco Antonio Zaccarini and Daniele Toti
https://ieeexplore.ieee.org/document/10296887

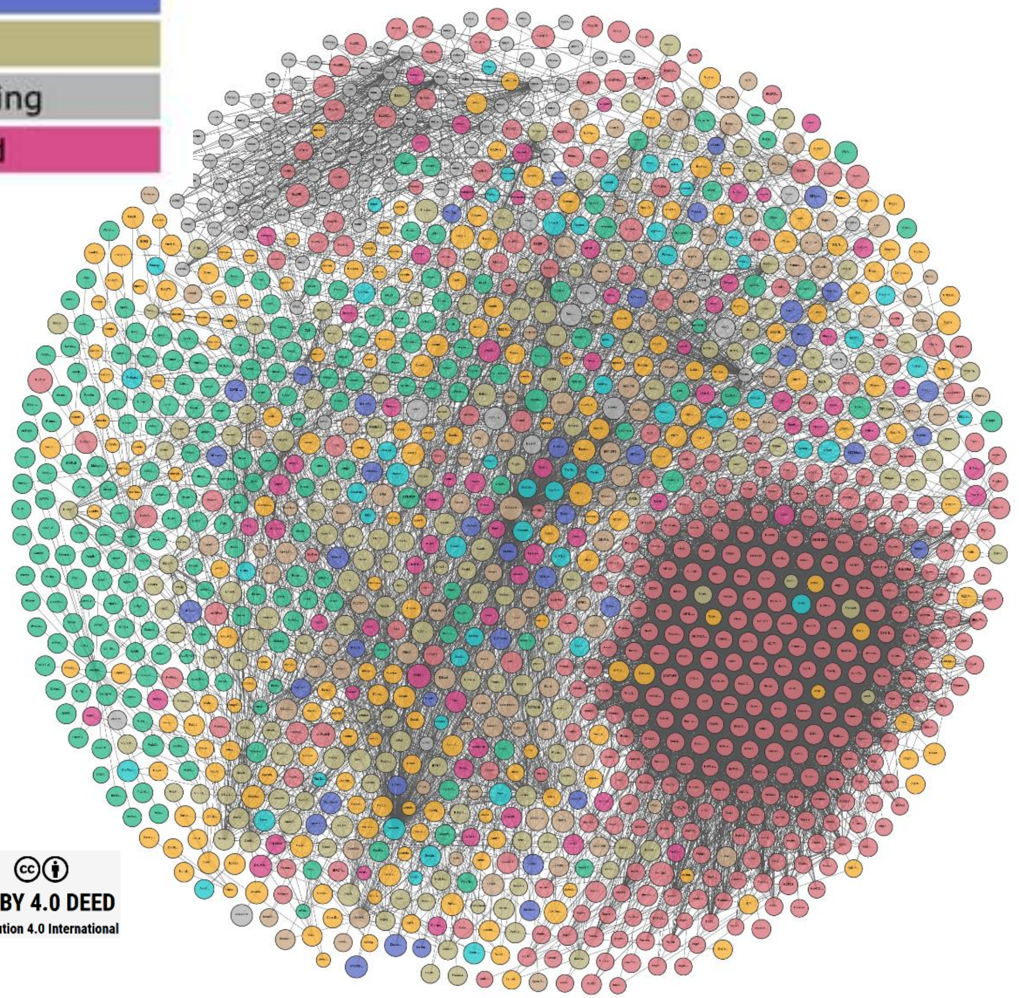
Linked Open Data Cloud

what do you notice?

Linked data is structured data which is interlinked with other data so it becomes more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers. Part of the vision of linked data is for the Internet to become a global database.[1]
Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term in a 2006 design note about the Semantic Web project.[2]
https://en.wikipedia.org/wiki/Linked_data

Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated



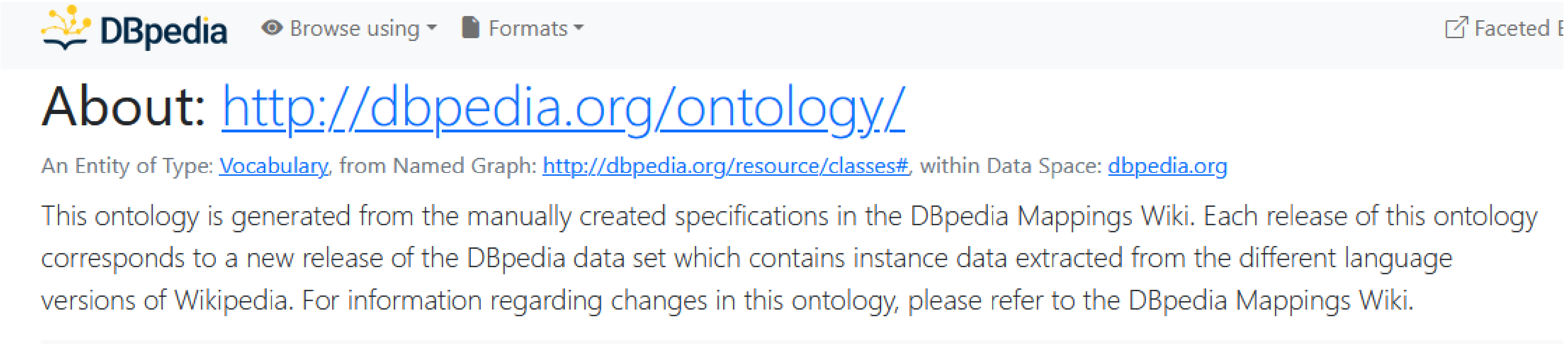
The dataset currently contains **1314** datasets with **16308** links

https://www.lod-cloud.net/

The Linked Open Data Cloud from lod-cloud.net

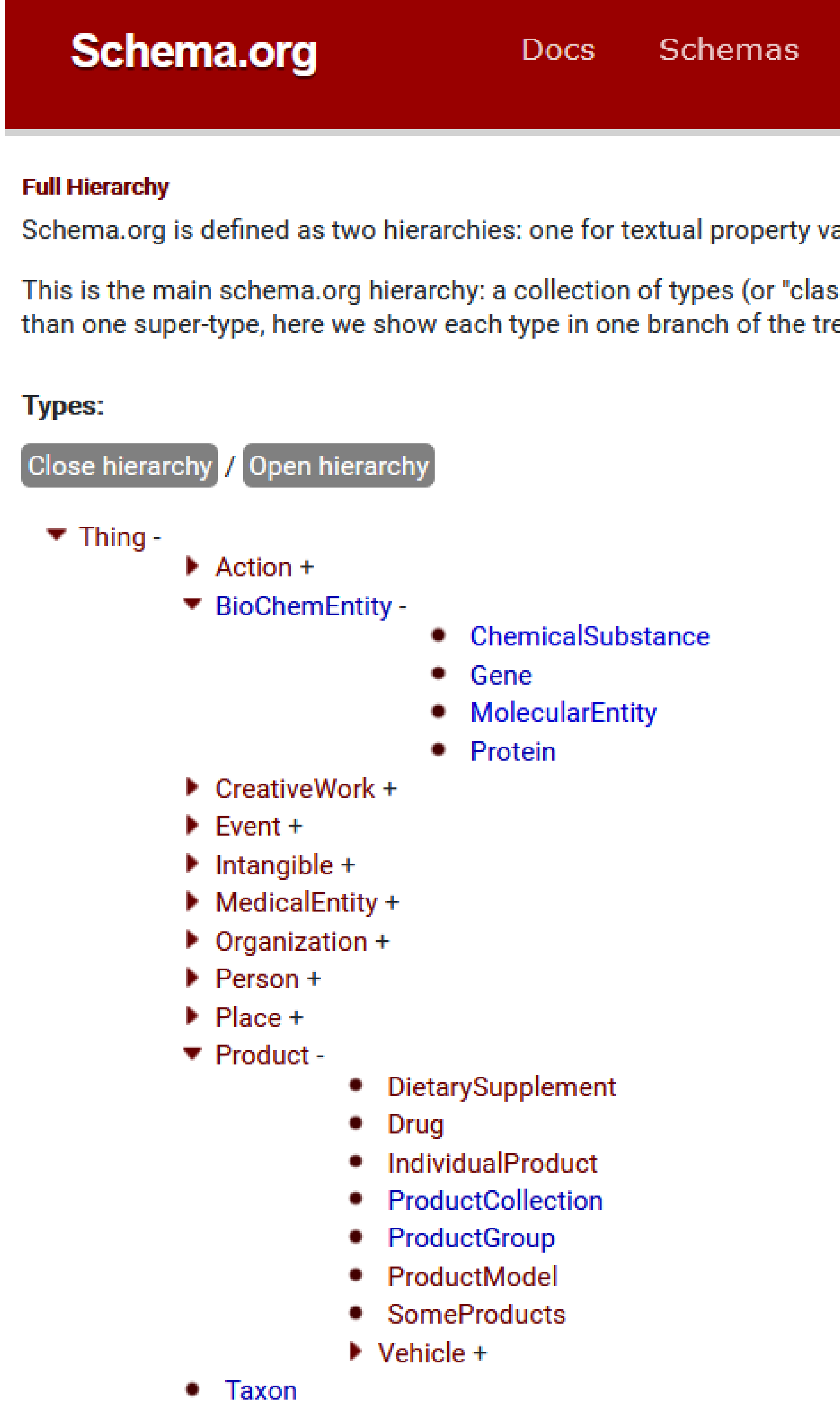
Schema.org
and also Dbpedia.org

- What do you notice?



- Same story: Where are materials?

Note: Dbpedia has Chemical Substance class (similar to Schema.org) but nothing about materials



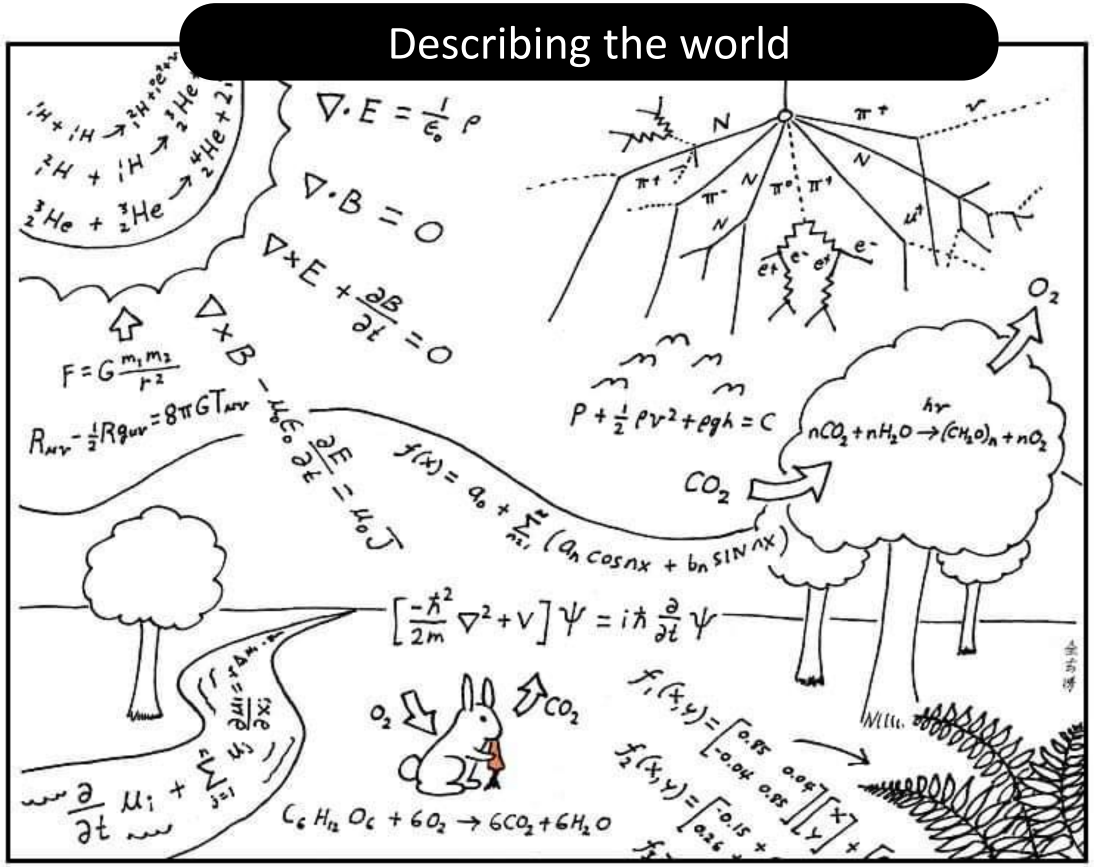
Challenges for semantic KM of Materials

- Heterogenous terminology is prevalent.
- EOSC Interoperability Framework presents the following challenges: https://eosc-portal.eu/eosc-interoperability-framework

- **Lack of common explicit definitions about the terms that are used by user communities**
- Lacking common semantic artefacts across communities (e.g., general ontologies that can be shared)

Elementary Multiperspective Material Ontology (EMMO)

- A multidisciplinary effort to develop a standard representational framework for applied sciences.
- Based on physics, analytical philosophy and information technologies.
- Pluralistic approach – the world can be described according to different perspectives



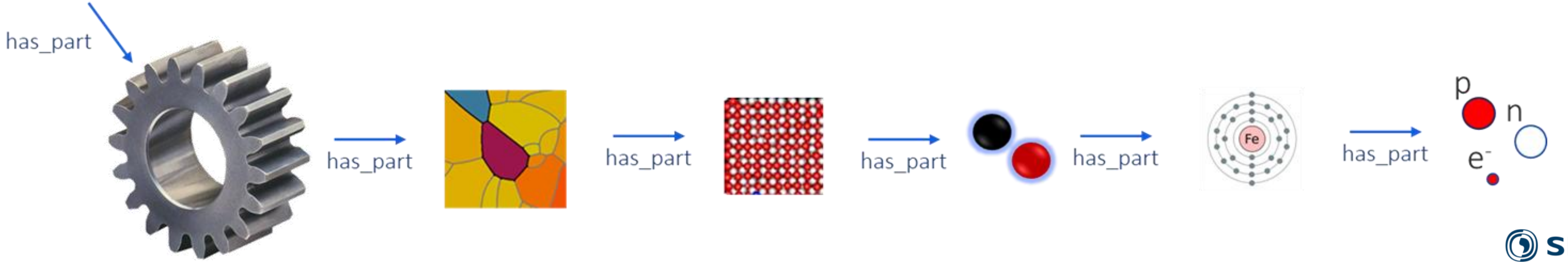
Creative Commons CC BY 4.0 license
Available on: https://github.com/emmo-repo/EMMO

SINTEF

Representing the World

The EMMO needs to represent **all the entities from sub-atomic to the universe**.

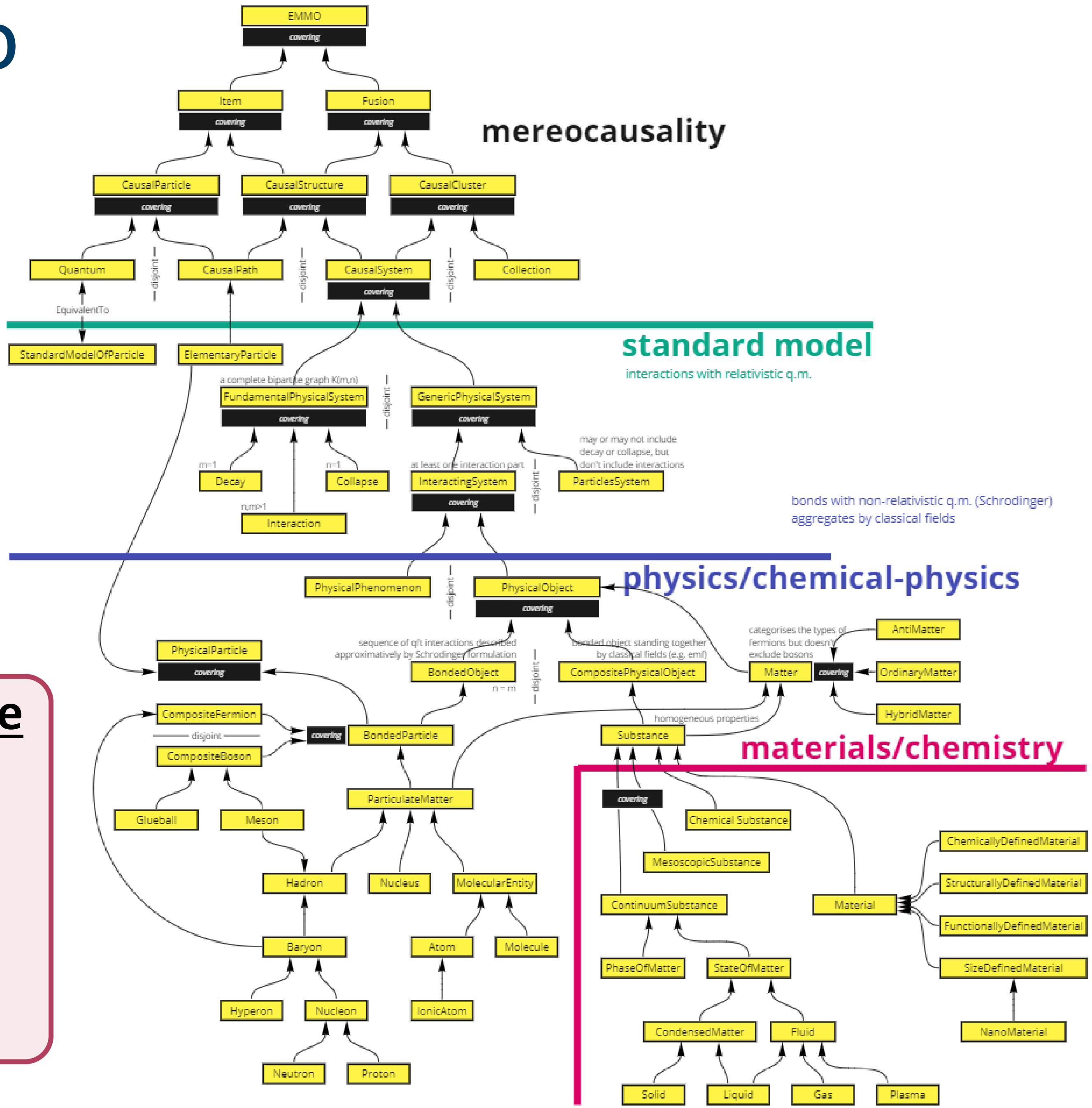
How do we deal with e.g., composite particles, atoms molecules, materials, phases, solids, mixtures?



From Elementaries to Materials

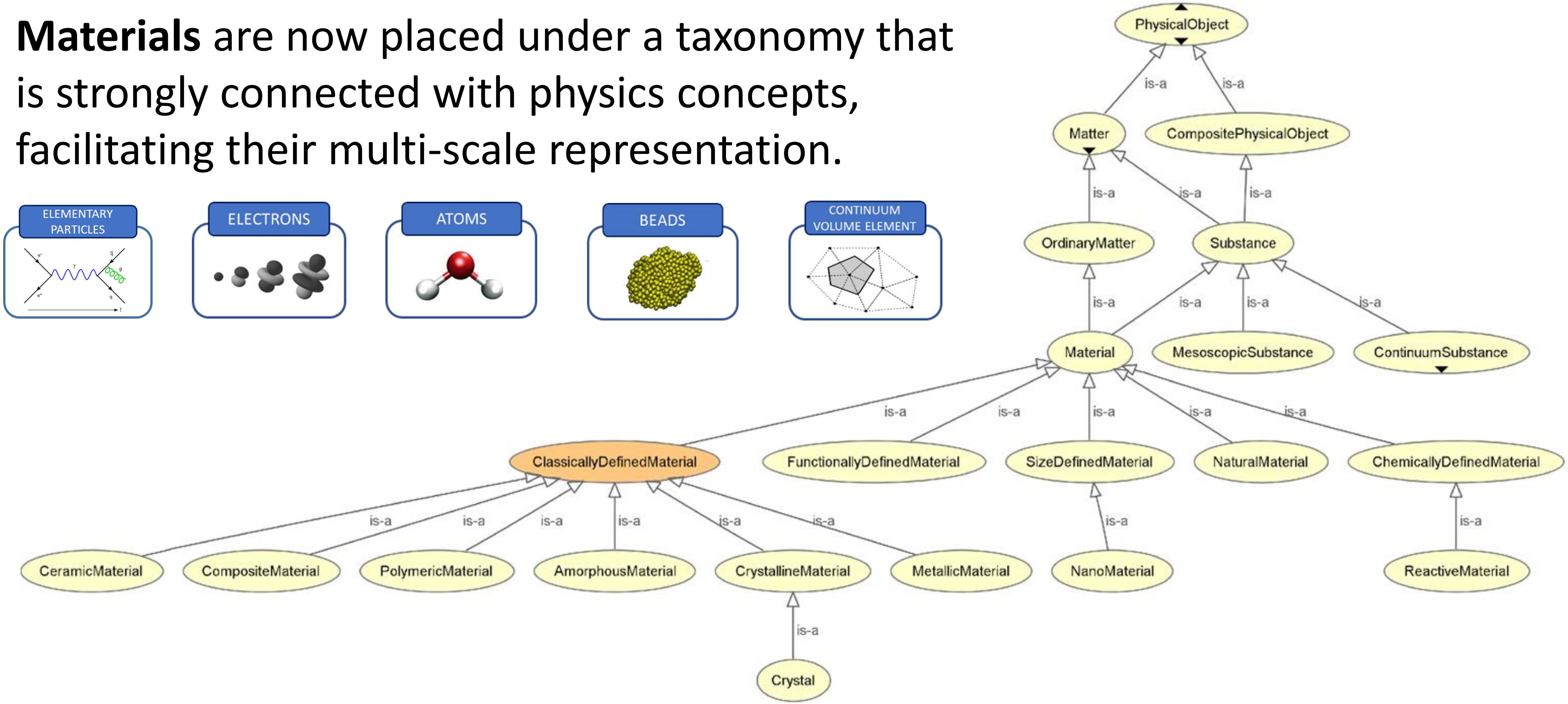
Seamlessly moving **from fundamental constituents to macroscopic objects**, including physical phenomena as described by physics.

EMMO 1.0.0 achieves <u>complete **multiscale** representational capabilities</u>, embracing and relating all entities from elementary particles to continuum, <u>including **physical models**</u> (e.g. QFT, Schrodinger, classical electrodynamics)

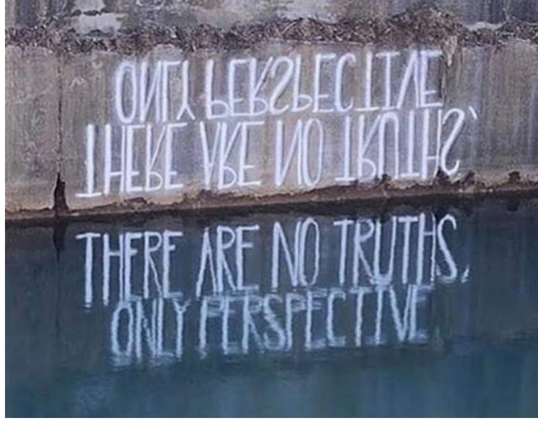


From Elementaries to Materials

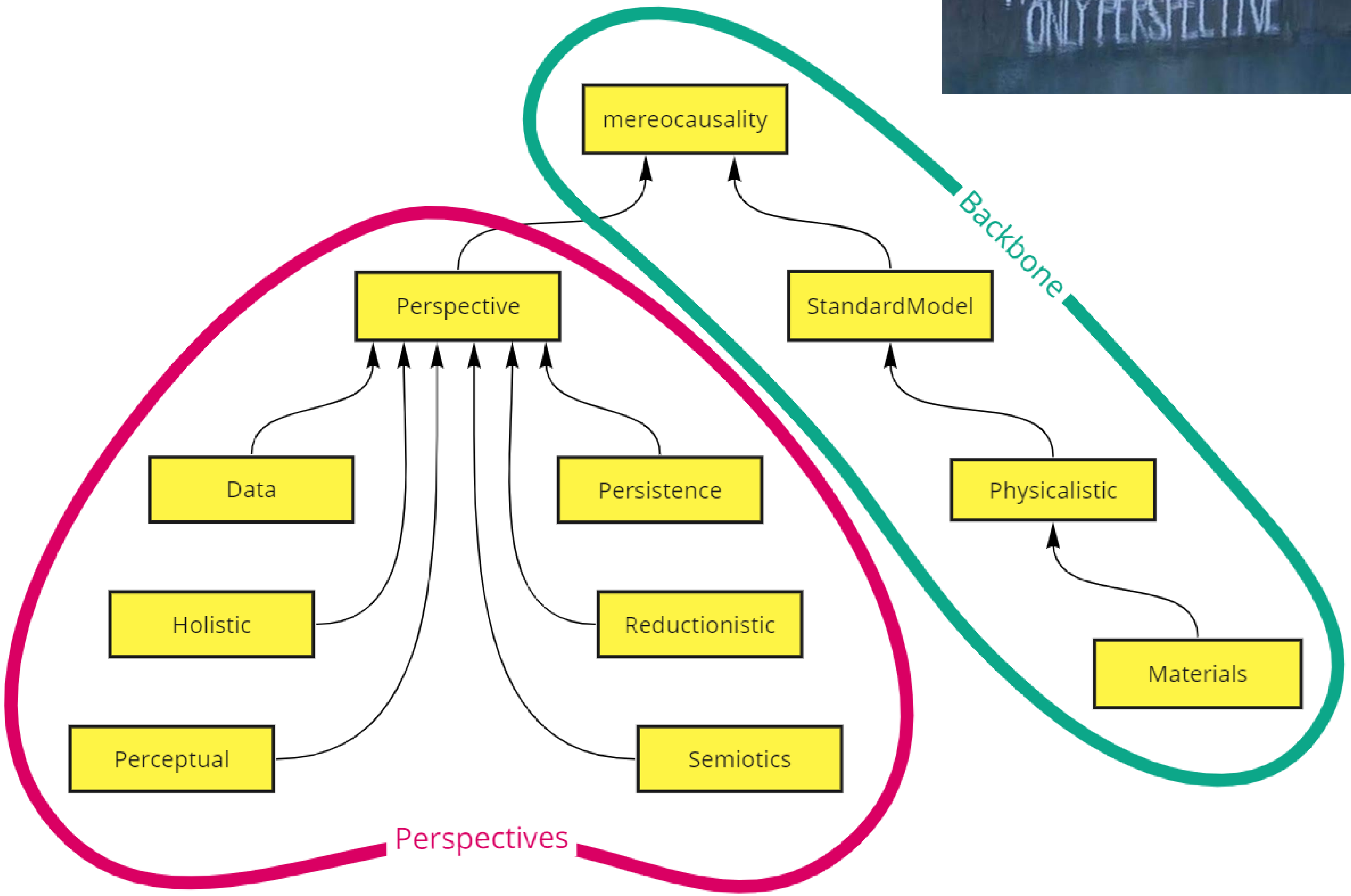
Materials are now placed under a taxonomy that is strongly connected with physics concepts, facilitating their multi-scale representation.



EMMO Perspectives



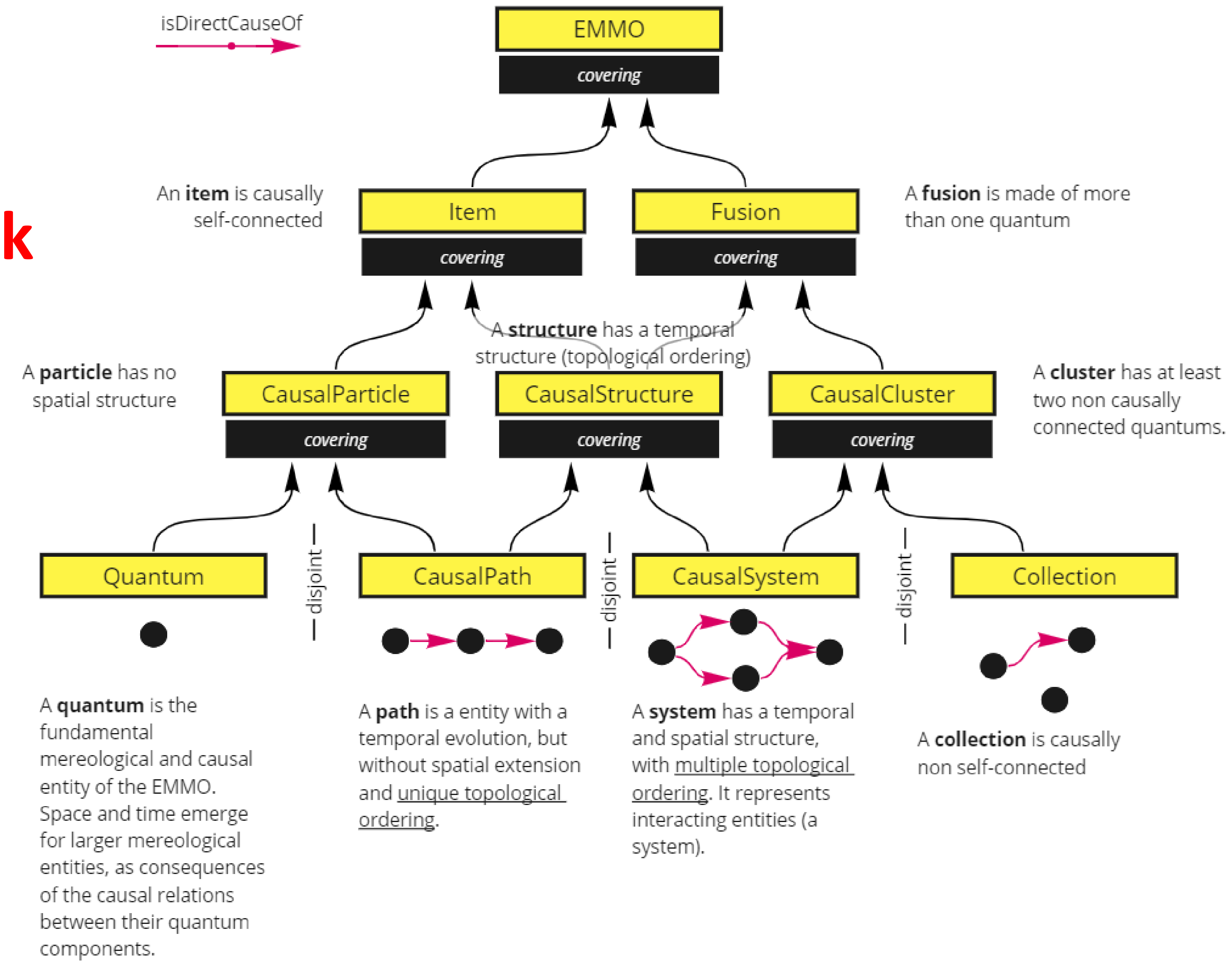
Clear separation between **physics-based representation modules** (quite complex and based on **applied sciences concepts**) and **generic representation modules** (perspectives, used to represent multi-domain and **common-sensical concepts**).



EMMO top Level

EMMO represent world entities as a **causal network** (a **Direct Acyclic Graph**) of **quantum entities** whose types and interactions are governed by the **Standard Model of Particles**.

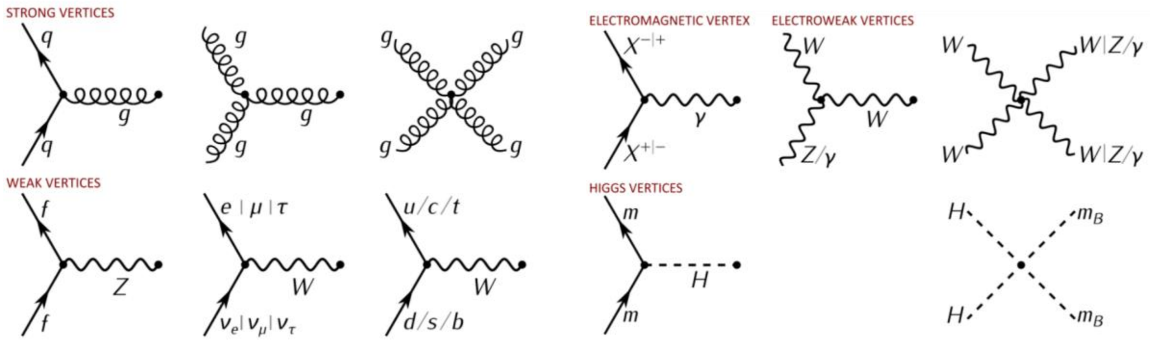
(Causality relations are invariants in general relativity)



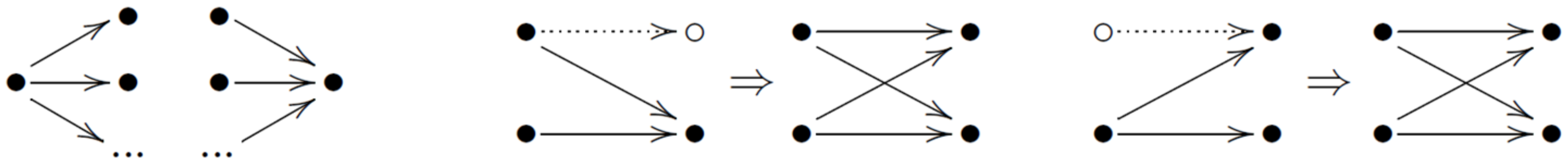
Radical Physicalism

By E. Ghedini et al.

Then we had to deal with **physical phenomena** describing how particles interacts.
We decided to build a causality theory to represent the **Quantum Field Theory** interactions, using **Feynman diagrams** representations.



It was amazing to see the theory grew almost naturally, and how easily the Feynman diagrams rules led to a **causal graph representation** of physical phenomena as **Direct Acyclic Graphs** with a **maximal graph dimension of 4**, totally in line with our **four-dimensional** approach.



(a) Minimal Causal Structure (b) Locality

Figure 1. EMMO's Core Naturalistic Commitments (the arrows stand for dC relations)

EMMO Mereocausality in First Order Logics

By E. Ghedini et al.

To go beyond the standard model, we decided to go for the fusion of **mereology** and **causality**, creating a new discipline called **mereocausality**, that has been formalised in First Order Logic.

The adoption of **AGEM** theory also led to the definition of **universe**: the **smallest** and the **largest** entities are described using these 13 axioms.

The EMMO mereocausality theory is **fully capable of representing everything that exist**… but is a bit too detailed for practical usage!!!

a1	$\mathtt{P}(x,x)$	*(Parthood: Reflexivity)*
a2	$\mathtt{P}(x,y) \land \mathtt{P}(y,x) \to x = y$	*(Parthood: Antisymmetry)*
a3	$\mathtt{P}(x,y) \land \mathtt{P}(y,z) \to \mathtt{P}(x,z)$	*(Parthood: Transitivity)*
a4	$\neg\mathtt{P}(y,x) \to \exists z(\mathtt{P}(z,y) \land \neg\mathtt{O}(z,x))$	*(Strong Supplementation)*
a5	$\exists x(\phi(x)) \to \exists z(\mathrm{F}\langle\phi(x)\rangle(z))$	*(Unrestricted Composition)*
a6	$\forall x \exists y(\mathtt{qP}(y,x))$	*(Atomicity)*
a7	$\neg\mathtt{C}(x,x)$	*(Causality: Irreflexivity)*
a8	$\mathtt{C}(x,y) \land \mathtt{C}(y,z) \to \mathtt{C}(x,z)$	*(Causality: Transitivity)*
a9	$\mathtt{C}(x,y) \to \mathtt{dC}(x,y) \lor \exists zw(\mathtt{C}(x,z) \land \mathtt{dC}(z,y) \land \mathtt{dC}(x,w) \land \mathtt{C}(w,y))$	*(Discreteness)*
a10	$\mathtt{C}(x,y) \to \mathrm{Q}(x) \land \mathrm{Q}(y)$	*(Quantum Causality)*
a11	$\mathtt{dC}(x,y) \to \exists z((\mathtt{dC}(x,z) \lor \mathtt{dC}(z,y)) \land y \neq z \land x \neq z)$	*(Minimal Causal Structure)*
a12	$\mathtt{dC}(x,y) \land \mathtt{dC}(x,z) \land \mathtt{dC}(w,y) \to \mathtt{dC}(w,z)$	*(Locality)*
a13	$\textsc{Item}(\mathbf{u})$	*(Connected Universe)*

$$\mathrm{qP}(x,y) \equiv \mathrm{P}(x,y) \land \mathbf{Q}(x) \qquad \mathrm{dC}(x,y) \equiv \mathrm{C}(x,y) \land \neg\exists z(\mathrm{C}(x,z) \land \mathrm{C}(z,y))$$

EMMO levels and landscape

Top Level
Mereo-causality
Standard Model
Perspectives (Data, Persistence, Holistic, Reductionistic, Semiotics,...)

Close to 1.0.0 release
Foundational paper upcoming

DCAT schema based EMMO ontology

Middle Level
Multi-perspective
Domain Reference/Discipline (Metrology, Models, Workflow, Substance, ...)

Major advance in metrology: Units of measure, Quantities including QUDT, logical consistency improvements

Domain Level
Domain Ontologies (Battery, Characterisation, Crystallography, ...)

Large effort in battery field. Used across EU battery projects. Fast growing, e.g. battery materials, connecting to CHAMEO characterisation ontology, Alignment with MDO, to connect to OptiMade databases ...

https://github.com/emmo-repo/EMMO

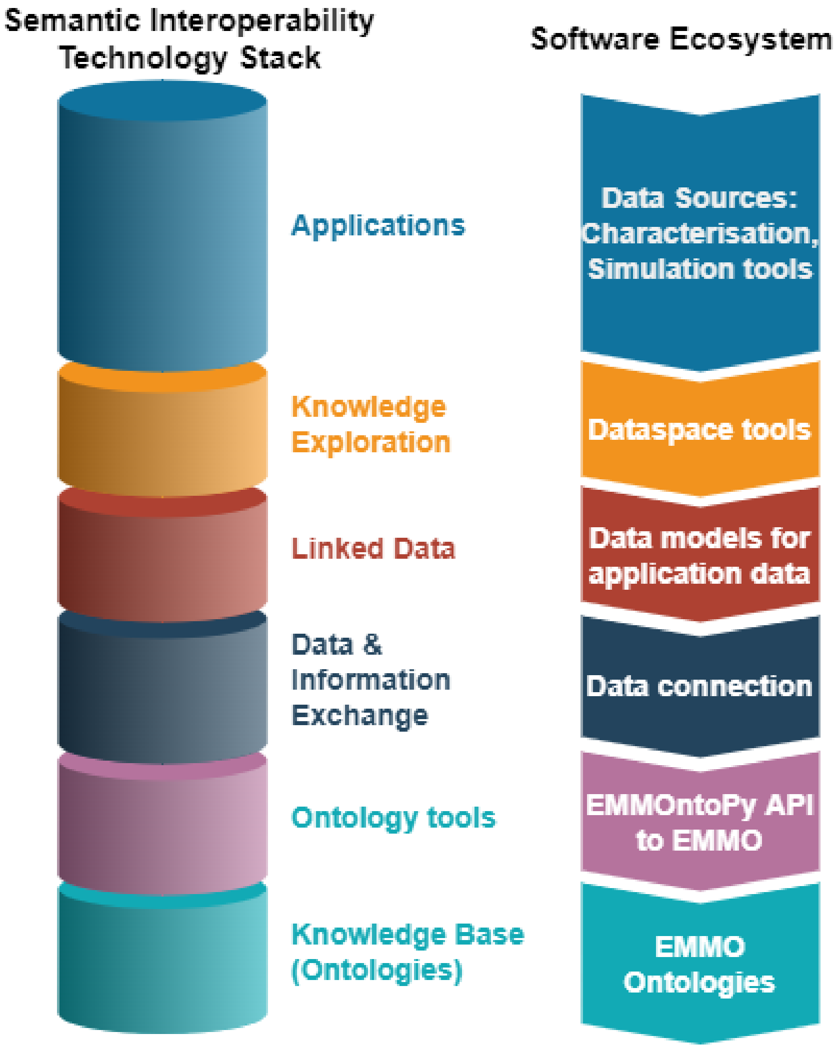
Battery Interface Genome – Materials Acceleration Platform
BIG-MAP

Vision of the EMMO ecosystem

EMMO Foundations	Disciplines	Domains	Applications
FOL, TLO	MLO, Metrology, Maths, Chemistry, Data etc	Battery, Characterisation etc	Battery Testing, R&D Data Integration
Univ Bologna and collaborators (CNR, …)	Unibo, GCL, SINTEF and collaborators from EU/National Projects	EU/National Projects, Industry Consortia	EU/National Projects, Commercial Projects
Logical consistency, rigour, achieving stable foundation, design principles, documentation, manage releases for TLO, MLO	Widely used concepts, use standards where possible, onion shell approach, consistency, tools for users	Generalisation of applications, Controlled Vocabularies, Domain object props, Domain namespaces	Application case development methodology, data connections, A-box, work with infrastructure and solutions providers
EMMC as useful over-arching legal body, but no legal function. Unibo responsible for TLO/MLO release	EMMO Editors Group (EMMC), coordinate project resources, responsible for releases of Discipline Module, Documentation, Tools	EMMC Task Groups for Domains, ensure stakeholders are involved. Bridging with non-EMMO DLOs, dissemination	Work with Domain TGs, feed requirements up the chain, Commercial Venture: Semantic Materials Ltd (?)

SINTEF

Technology stack



DLite https://github.com/SINTEF/dlite

Tripper https://github.com/EMMC-ASBL/tripper

https://github.com/emmo-repo/EMMO-python

https://github.com/emmo-repo/EMMO

SINTEF

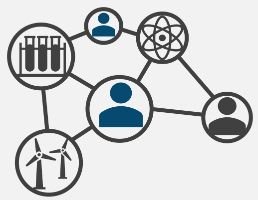
Implementation

Levels of data documentation

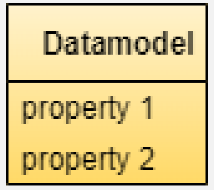
… and what aspects of FAIR data they address

FAIR **Cataloguing**
Where and how datasets can be found and accessed

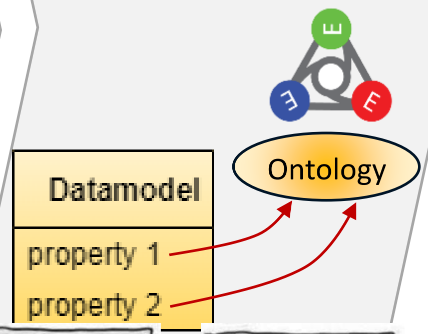
FAIR **Contextual documentation**
How datasets relate to other data and things (linked data)

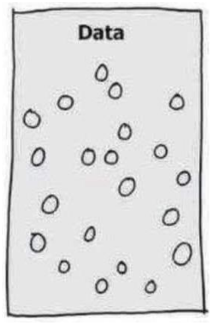


FAIR **Structural documentation**
How datasets are structured and represented numerically

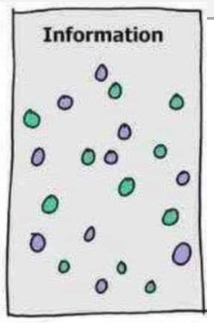


FAIR **Semantic documentation**
What is the meaning of the data

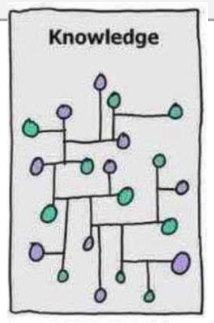




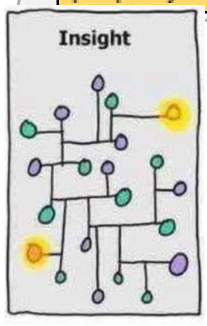
symbols



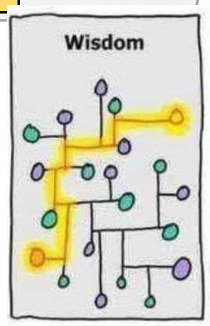
who, what, where…



how



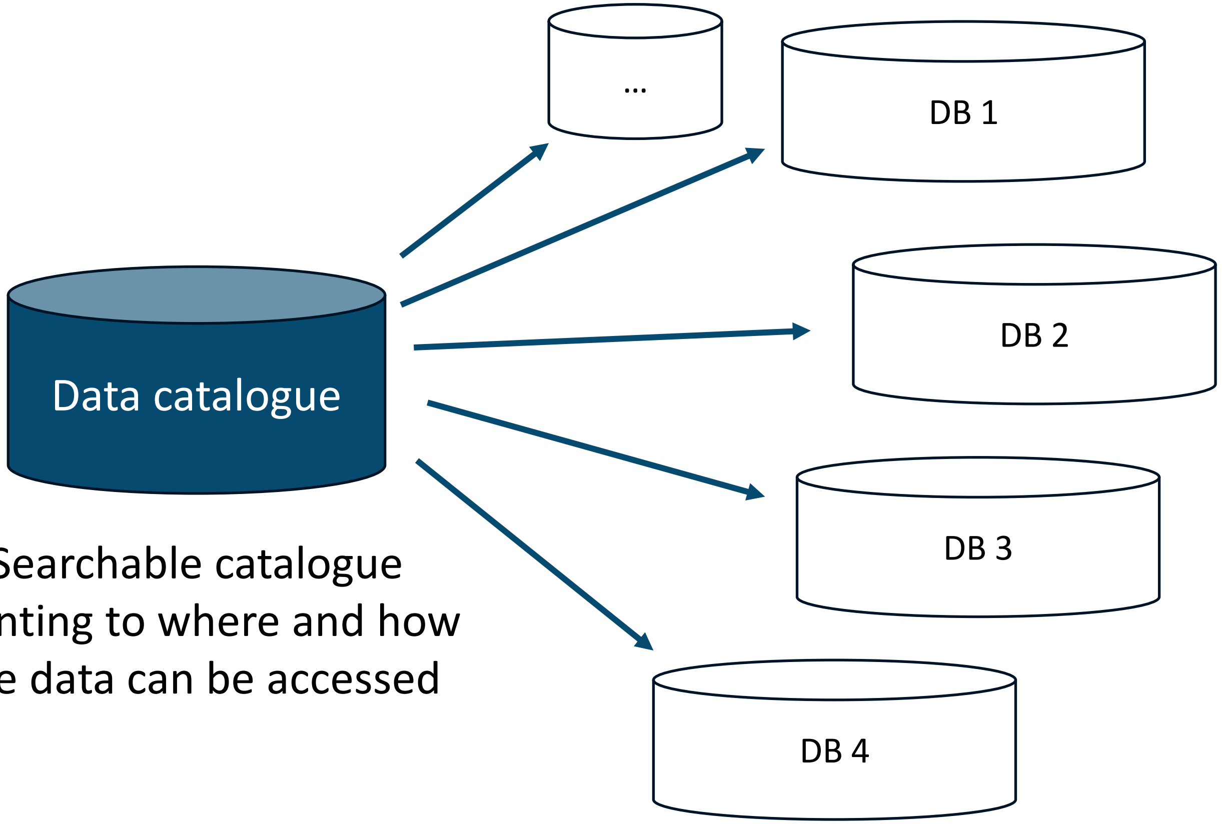
why



evaluated insight

Cataloguing

How can existing and new data be found and accessed



Searchable catalogue pointing to where and how the data can be accessed

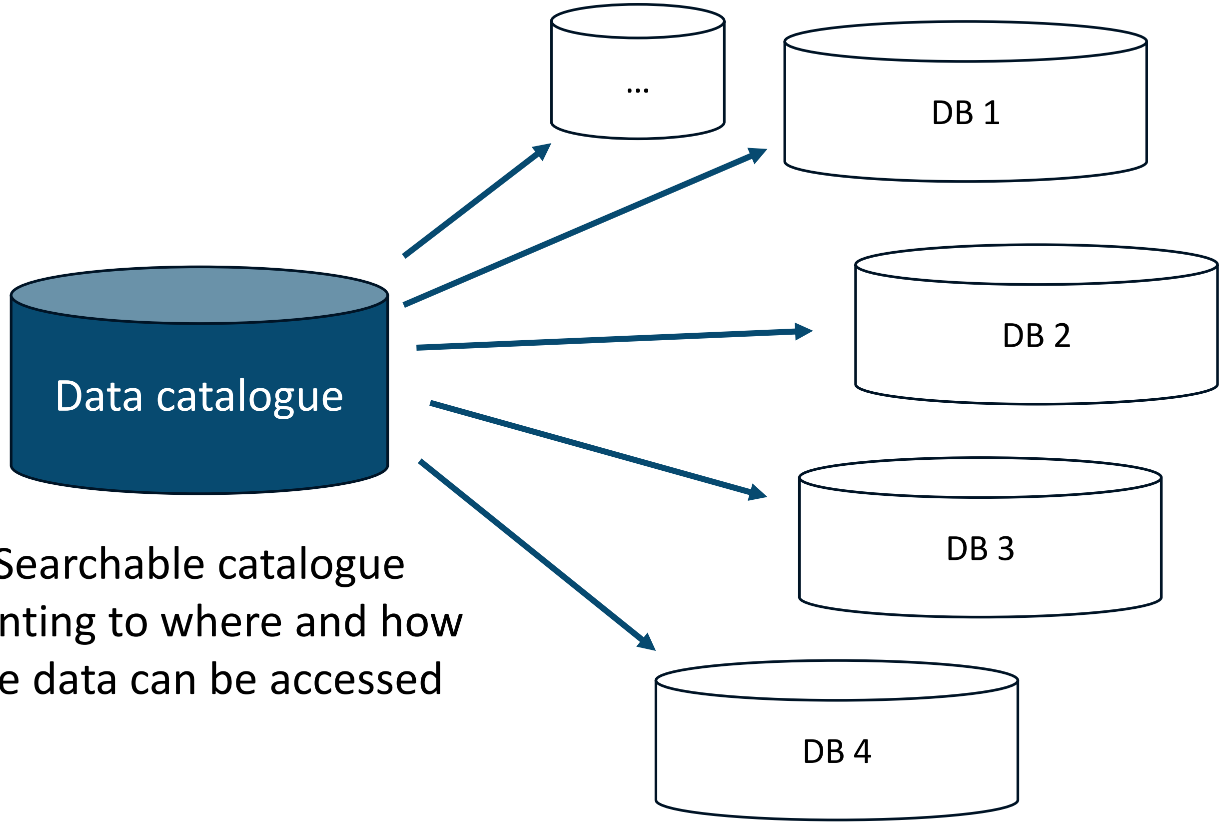
Backend databases and file archives with the actual data

Cataloguing

How can existing and new data be found and accessed

W3C standard for data catalogues

Further specialised by the European DCAT application profile (DCAT-AP) and the Norwegian DCAT-AP-NO



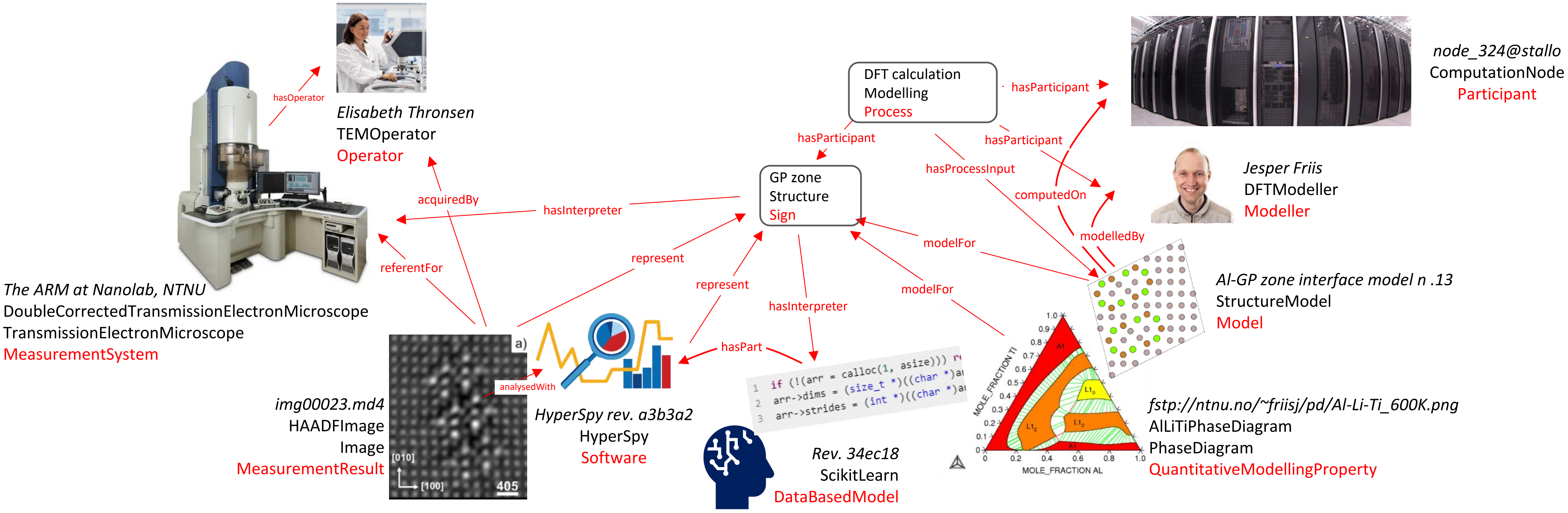
Searchable catalogue pointing to where and how the data can be accessed

Backend databases and file archives with the actual data

SINTEF

Contextual documentation

How the datasets relates to other datasets – linked data

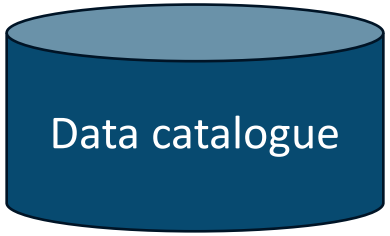


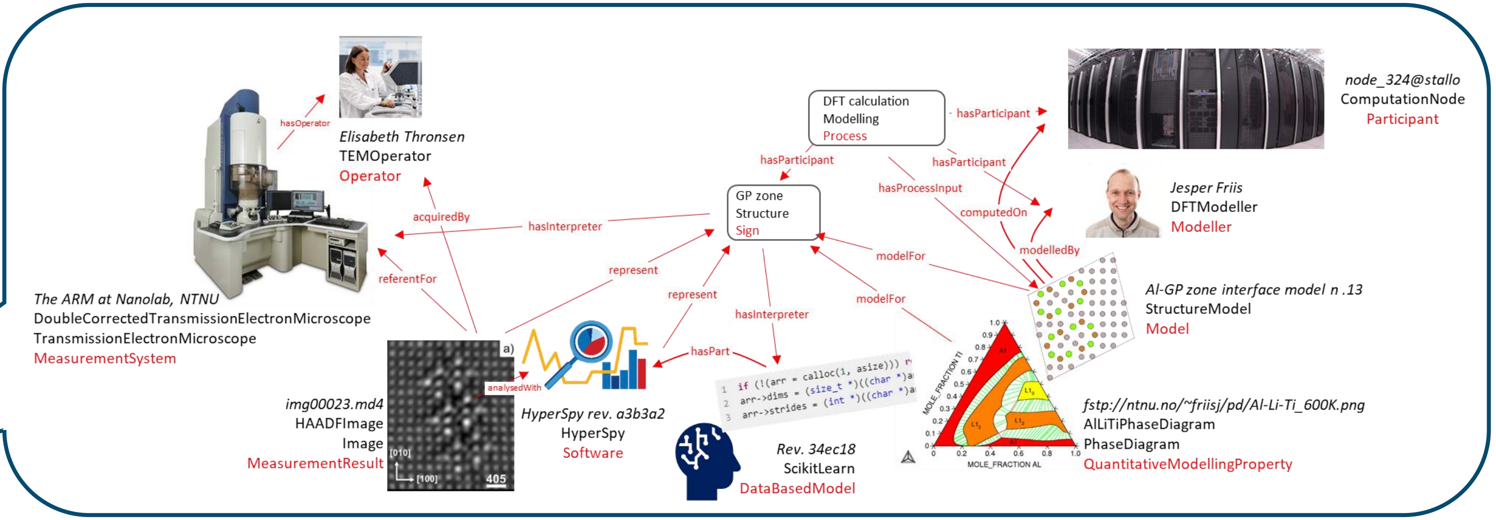
SINTEF

Contextual documentation

How the datasets relates to other datasets – linked data

Reusability
Traceability





knowledge graph

- If we want to be able to explore the relations between datasets, the data catalogue should contain the knowledge graph.
- Graph databases (triplestores) are designed for that.
- DCAT – a W3C standard (vocabulary) for data catalogues.
- SPARQL – query language for knowledge graphs

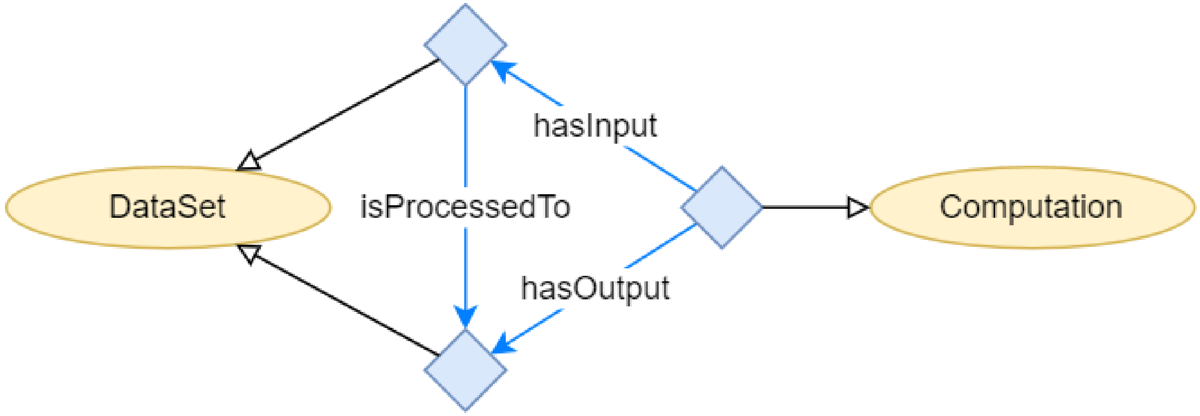
SINTEF

Contextual documentation

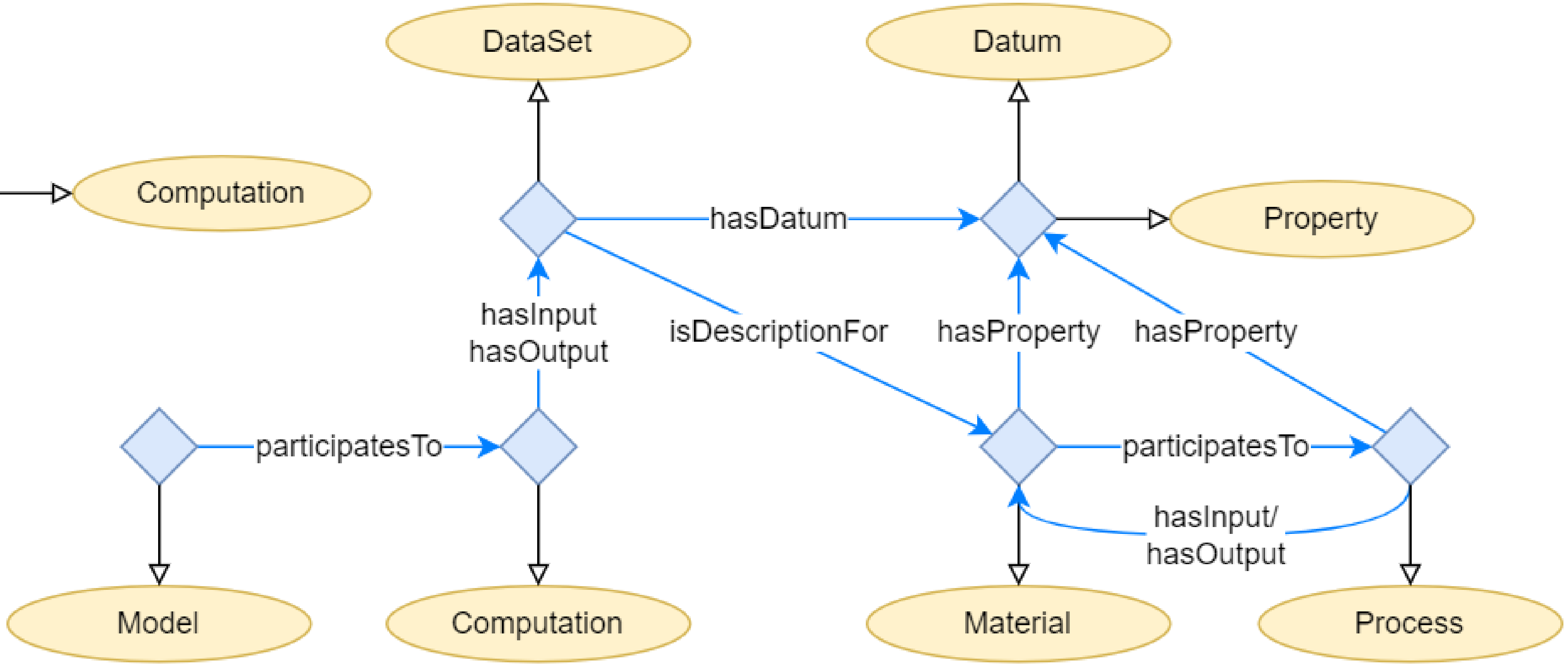
Extended semantics with the palette of basic relations between datasets, materials, properties, models, …

Some examples

Traceability



Reusability (meaning)



SINTEF

Structural documentation

How data is structured numerically

Datamodel	
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage
Description:	SEM image with elemental mappings. Represented as…

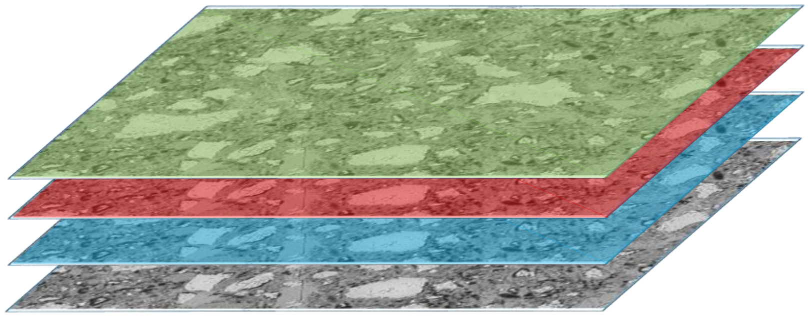
Dimensions

Name	Description
nelements	Number of elements for elemental mapping.
height	Number of pixels along image height.
width	Number of pixels along image width.

Properties

Name	Type	Shape	Unit	Description
data	float32	[nelements+1, height, width]		Image data
elements	string	[nelements]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref http://.../0.1/SEM			Ref. to data model for SEM metadata

Simple example – SEM image with elemental mapping



Data cube

Structural documentation

How data is structured numerically

Datamodel	
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage
Description:	SEM image with elemental mappings. Represented as…

Dimensions

Name	Description
nelements	Number of elements for elemental mapping.
height	Number of pixels along image height.
width	Number of pixels along image width.

Properties

Name	Type	Shape	Unit	Description
data	float32	[nelements, height, width]		Image data
elements	string	[nelements]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref			Ref. to data model for SEM metadata

Serialised as YAML


```yaml
uri: http://onto-ns.com/meta/matchmaker/0.1/SEMImage
description: |
  SEM image with elemental mappings. Represented as a stack of elemental mapping
  followed by the image formed by the back-scattered electrons (BSE).
  Set `nelements=0` if you only have the back-scattered image.
dimensions:
  nelements: Number of elements for elemental mapping.
  height: Number of pixels along the image height.
  width: Number of pixels along the image width.
properties:
  data:
    type: float32
    shape: [nelements+1, height, width]
    description: "Image data - a stack of images for each channel"
  elements:
    type: string
    shape: [nelements]
    description: "The chemical symbols of the element maps."
  pixelwidth:
    type: float64
    unit: m
    description: "Width of each pixel."
  pixelheight:
    type: float64
    unit: m
    description: "Height of each pixel."
  metadata:
    type: ref
    ref: http://onto-ns.com/meta/characterisation/SEM/0.1/SEM
    description: "Reference to data model for SEM metadata."
```


SINTEF

Structural documentation

How data is structured numerically

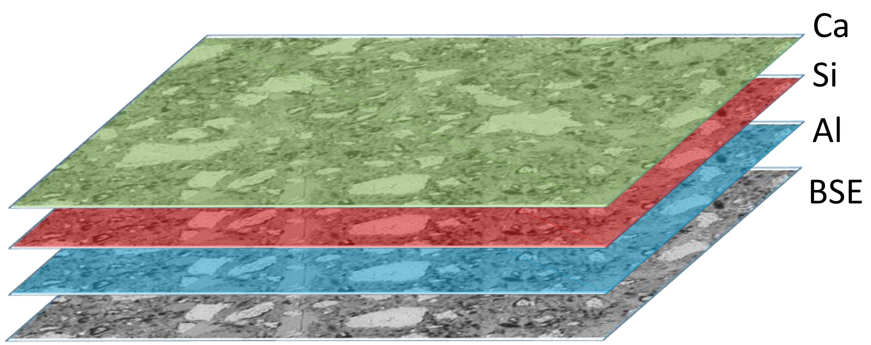
Datamodel	
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage
Description:	SEM image with elemental mappings. Represented as…

Dimensions	
Name	**Description**
nelements	Number of elements for elemental mapping.
height	Number of pixels along image height.
width	Number of pixels along image width.

Properties				
Name	**Type**	**Shape**	**Unit**	**Description**
data	float32	[nelements+1, height, width]		Image data
elements	string	[nelements+1]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref http://…/0.1/SEM			Ref. to data model for SEM metadata

Instance


```
{
  "88ca46ff-8404-48d9-b4b9-2140c6b3bdff": {
    "meta": "http://onto-ns.com/meta/matchmaker/0.1/SIMImage",
    "dimensions": [3, 2560, 1920],
    "properties": {
      "data": [[[…], …], …],
      "elements": ["Ca", "Si", "Al", "BSE"],
      "pixelwidth": 0.5e-6,
      "pixelheight": 0.5e-6,
      "metadata": "dcf0d99a-ef85-4efe-afee-bd57fef5ad5f"
    }
  }
}
```

Data cube

SINTEF

Semantic documentation

What is the meaning of the data – map data model properties to ontologies

Datamodel	
Identity (uri):	http://onto-ns.com/meta/matchmaker/0.1/SEMImage
Description:	SEM image with elemental mappings. Represented as…

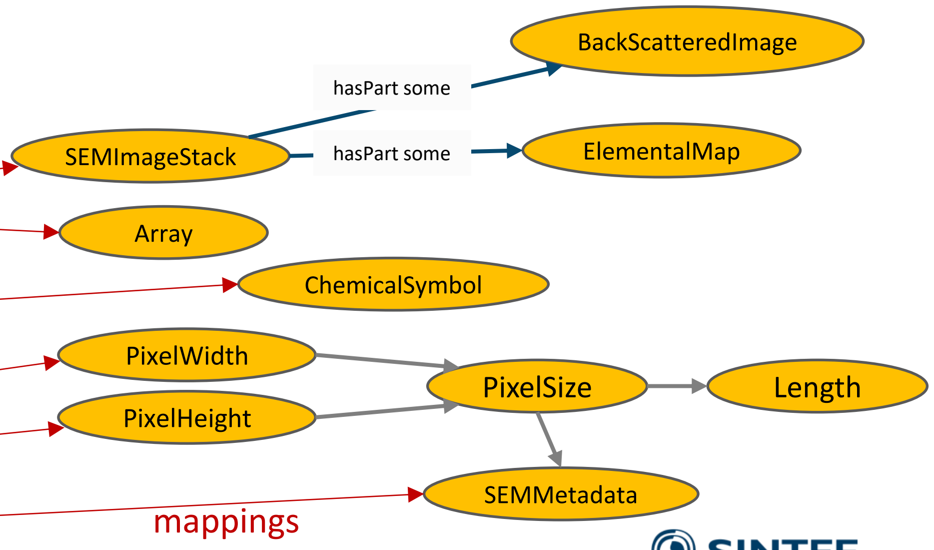
Dimensions

Name	Description
nelements	Number of elements for elemental mapping.
height	Number of pixels along image height.
width	Number of pixels along image width.

Properties

Name	Type	Shape	Unit	Description
data	float32	[nelements, height, width]		Image data
elements	string	[nelements]		Chemical symbols of elements for elemental mapping
pixelwidth	float64		m	Width of each pixel
pixelheight	float64		m	Height of each pixel
metadata	ref			Ref. to data model for SEM metadata

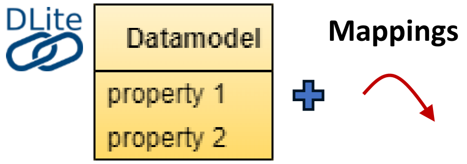
Ontologies
Provide shared meaning



SINTEF

Semantic documentation: structure & meaning

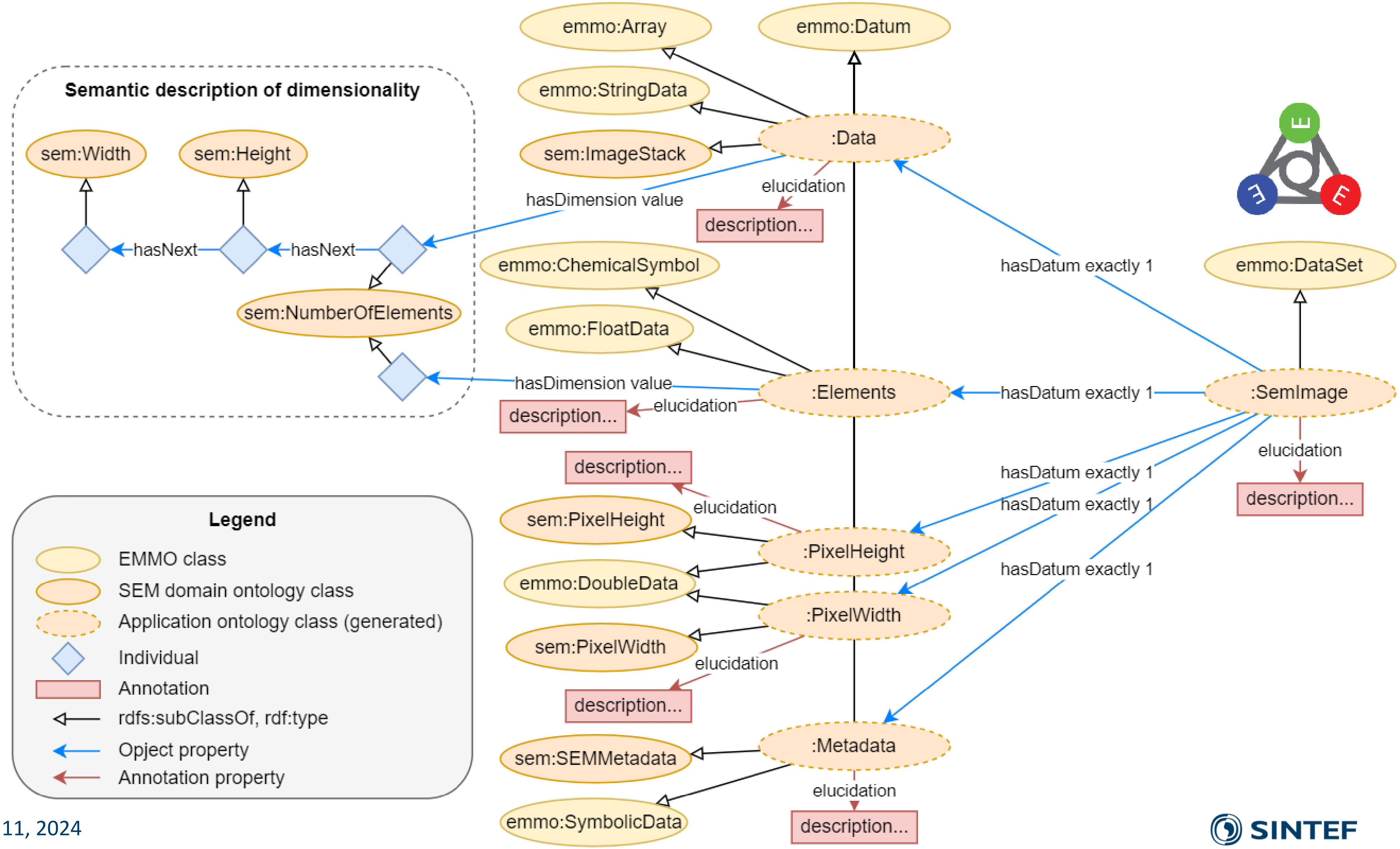
Equivalent to



or

Any other interopera-bility system under-standing EMMO

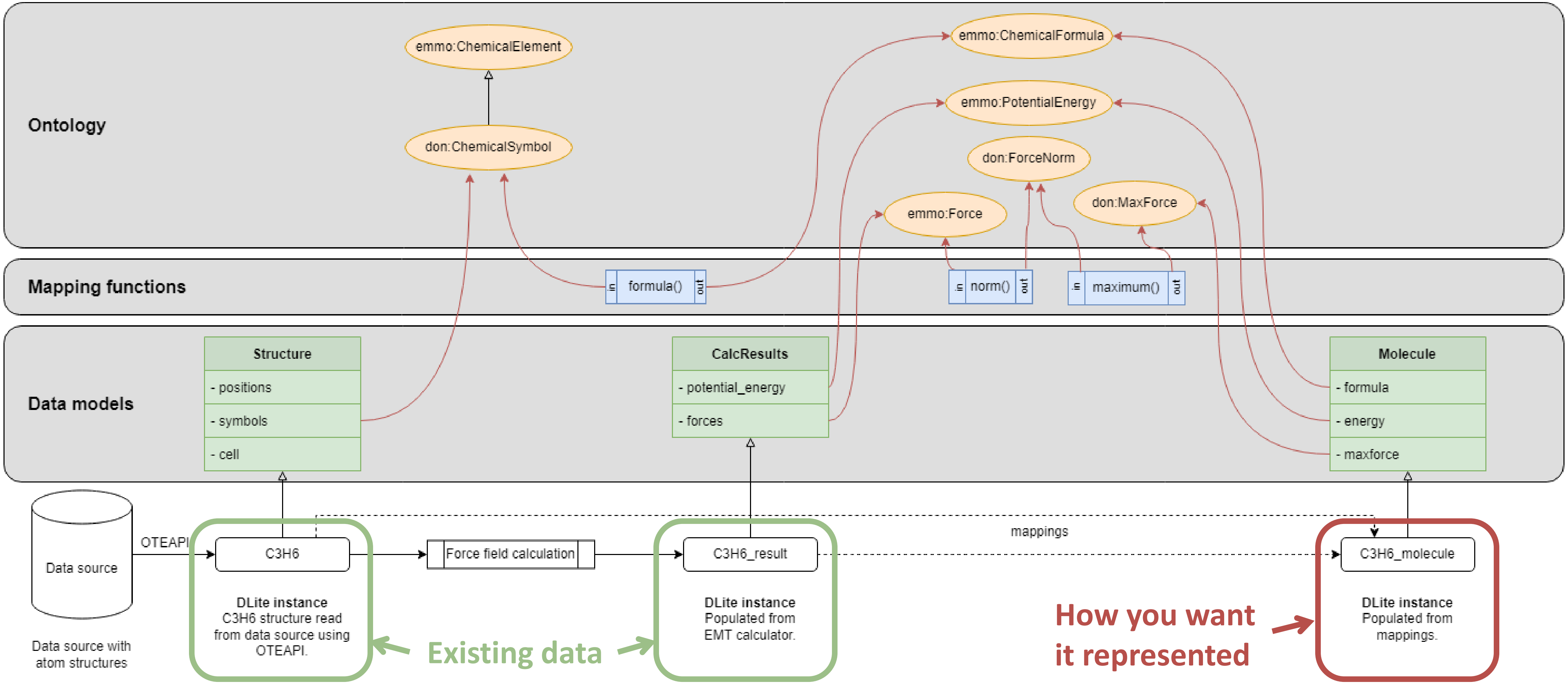
Common language for semantic data exchange



SINTEF

Semantic interoperability

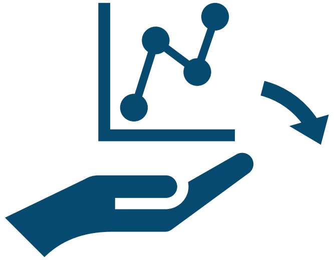
Get the data in the form you want

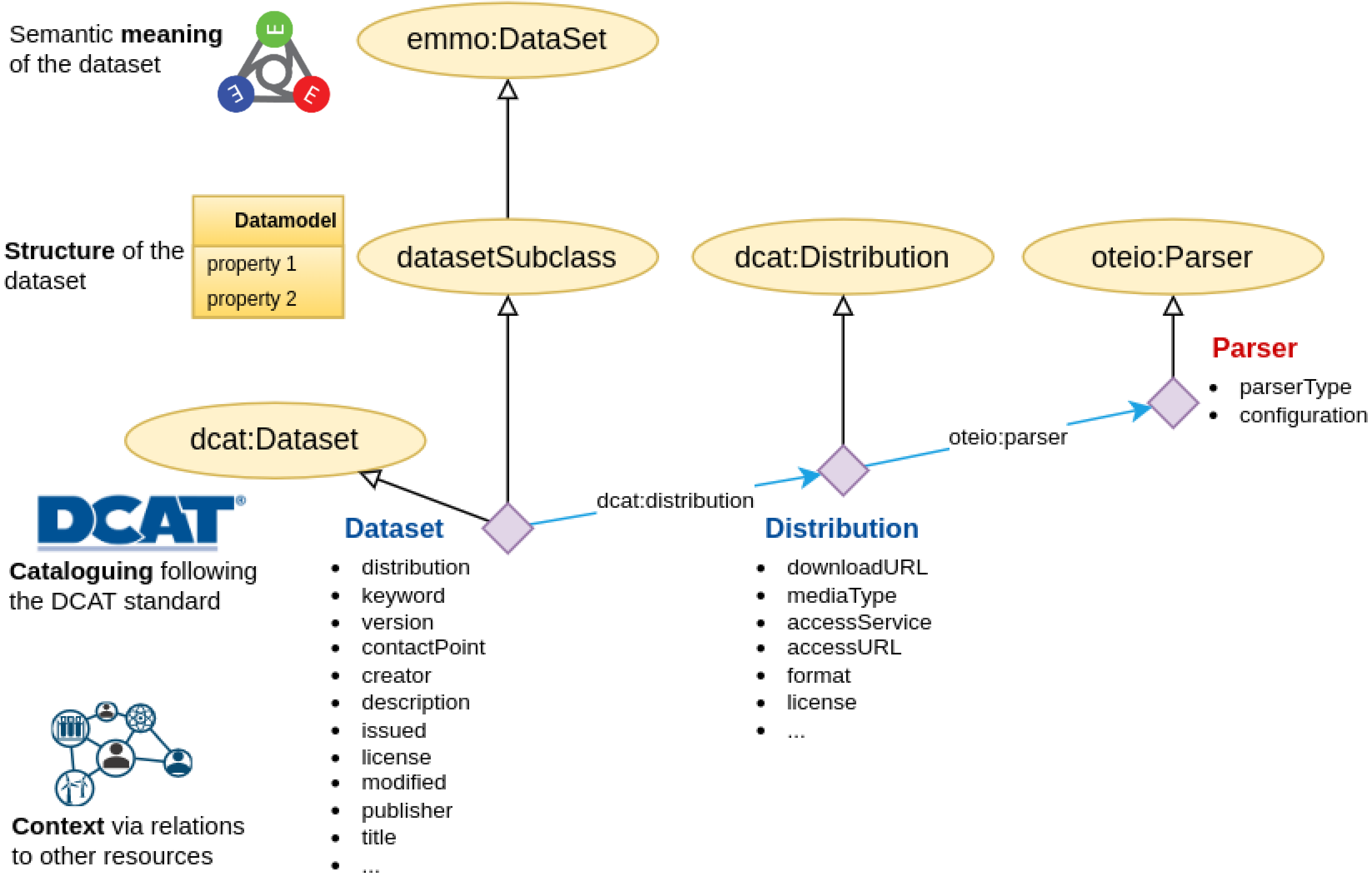


SINTEF

Documenting a data source

Data provider





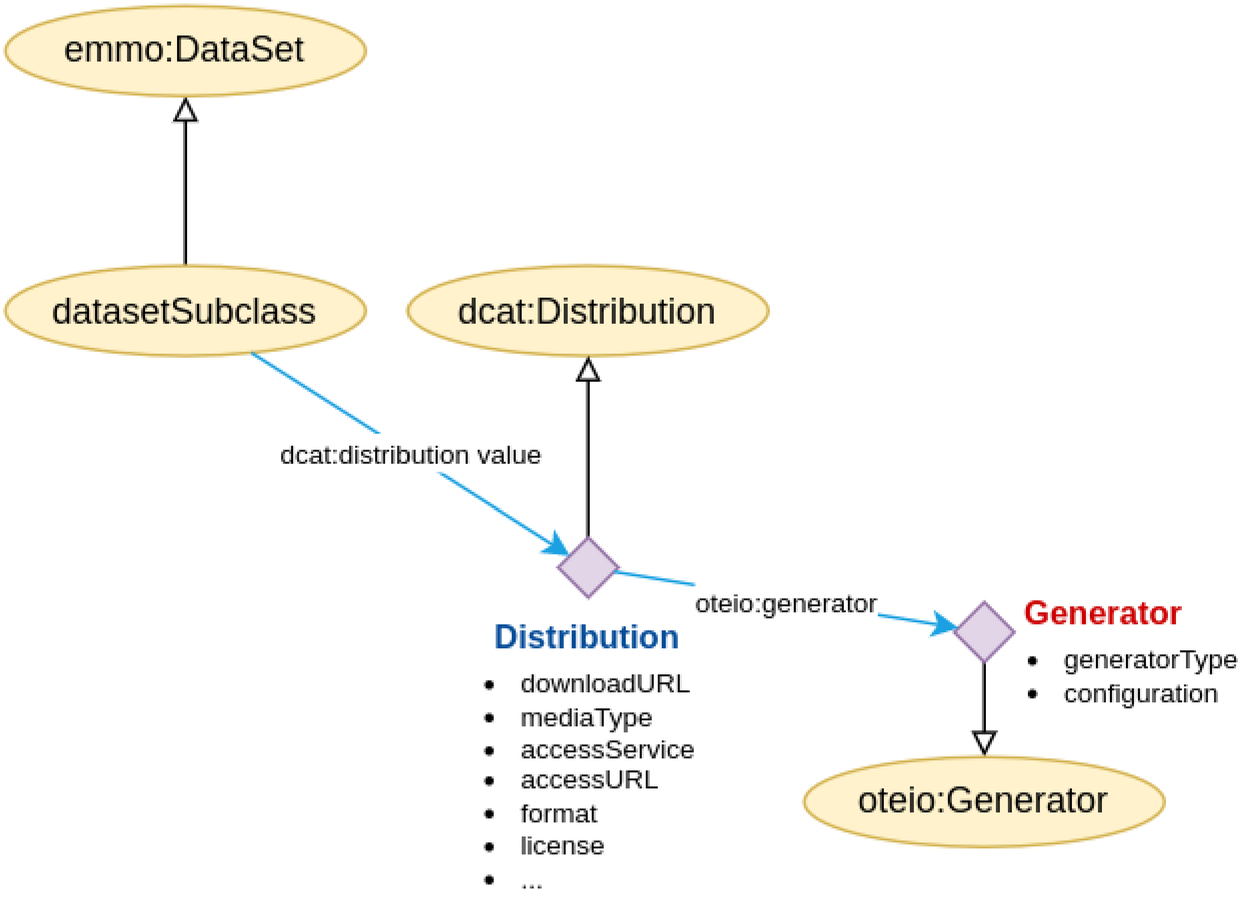
Documenting a data sink

Data consumer

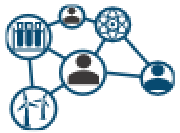
Semantic **meaning** of the dataset

Structure of the dataset

Datamodel
property 1
property 2



Cataloguing following the DCAT standard



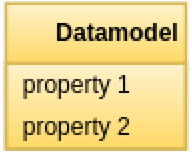
Context via relations to other resources

Documenting a data sink

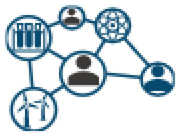
Data consumer

Semantic **meaning** of the dataset

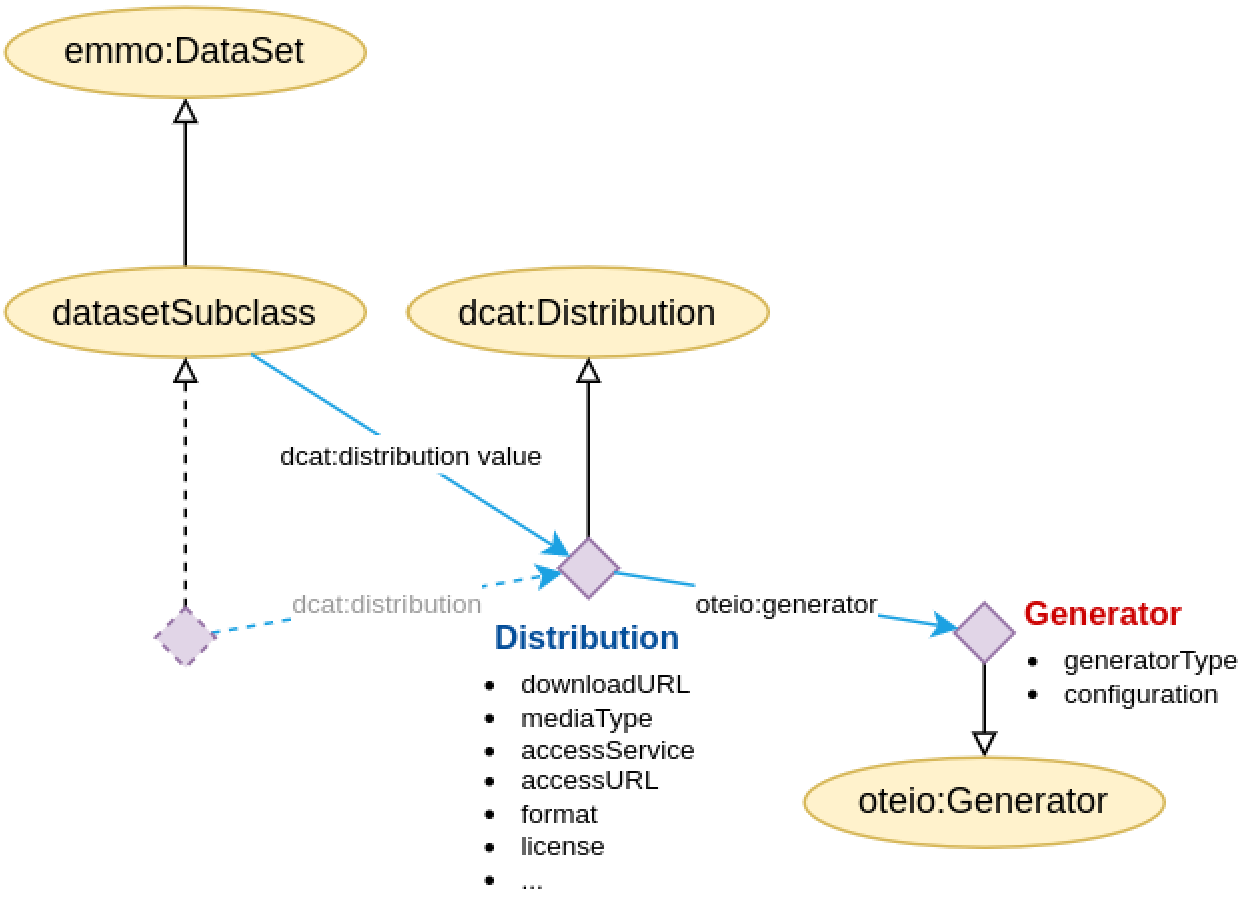
Structure of the dataset



Cataloguing following the DCAT standard

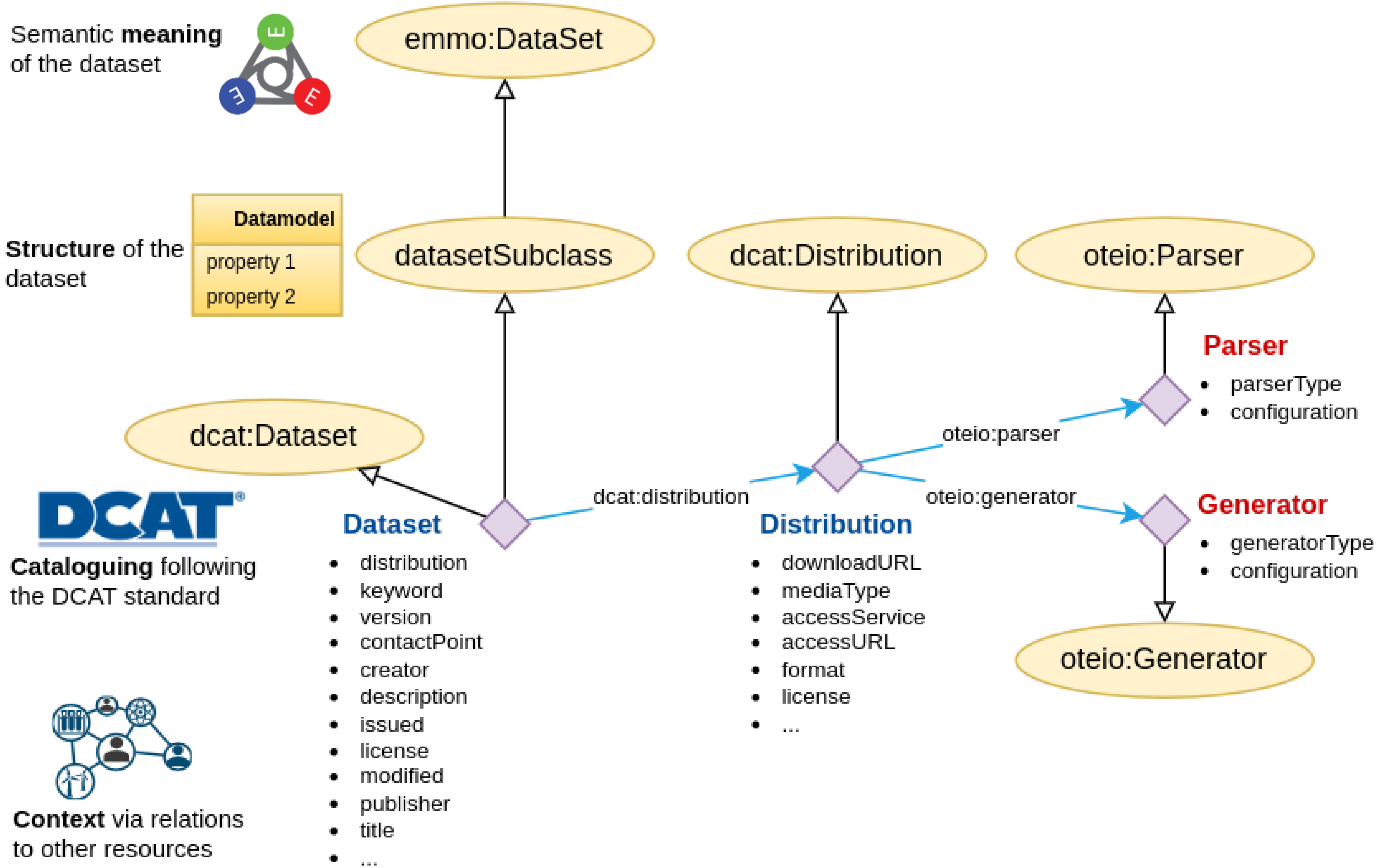


Context via relations to other resources



SINTEF

Data documentation



SINTEF

Thanks

Acknowledgement:

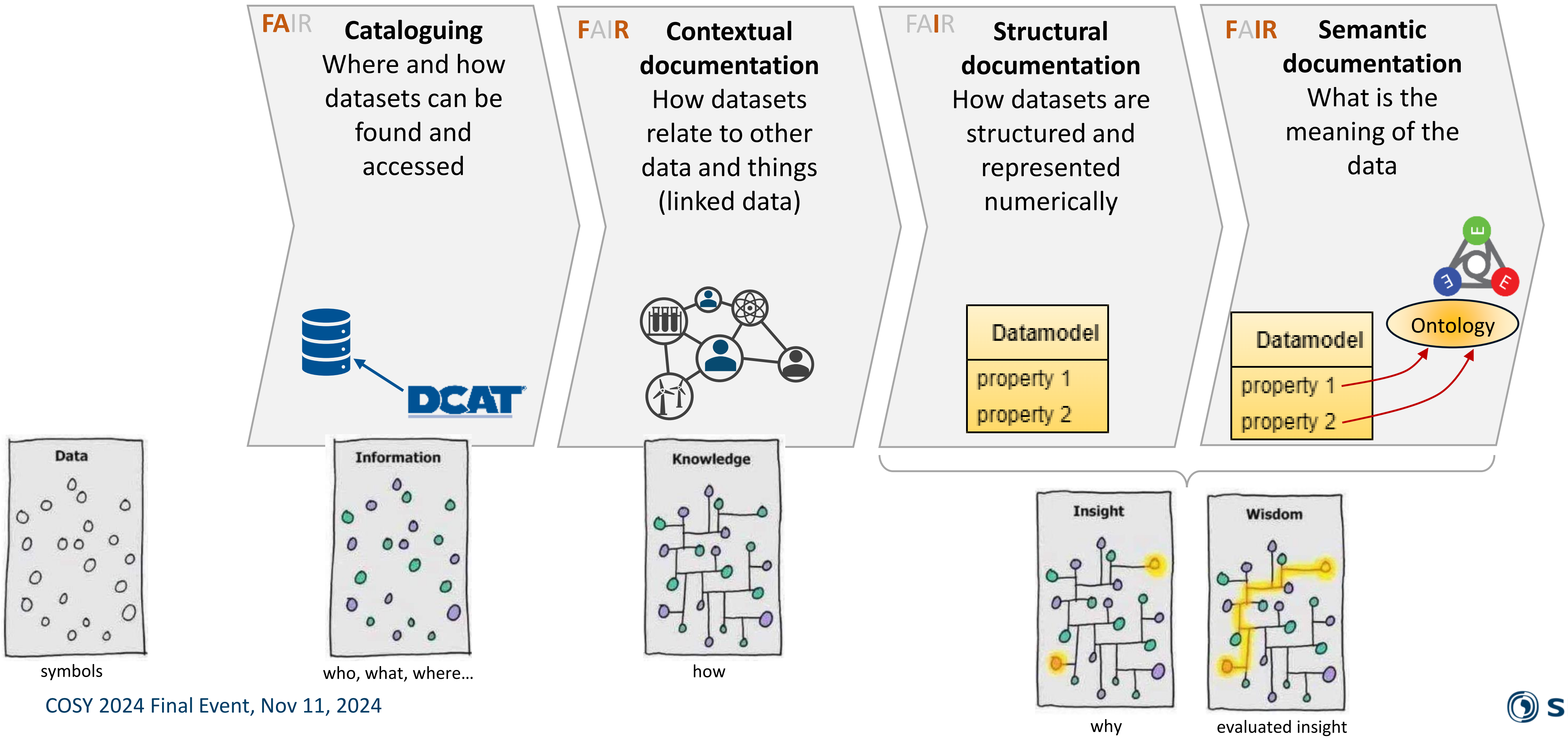
This work has received funding from the *European Union's Horizon 2020 research and innovation programme* and Norwegian Research Council via the following projects:

- OntoTrans (2020-2024) Grant Agreement n. 862136
- SFI PhysMet (2020-2028) Research Council of Norway n. 309584
- OpenModel (2021-2025) Grant Agreement n. 953167
- NanoMECommons (2021-2025) Grant Agreement n. 952869
- MatCHMaker (2022-2026) Grant Agreement n. 101091687
- PINK (2024-2028) Grant Agreement n. 101137809

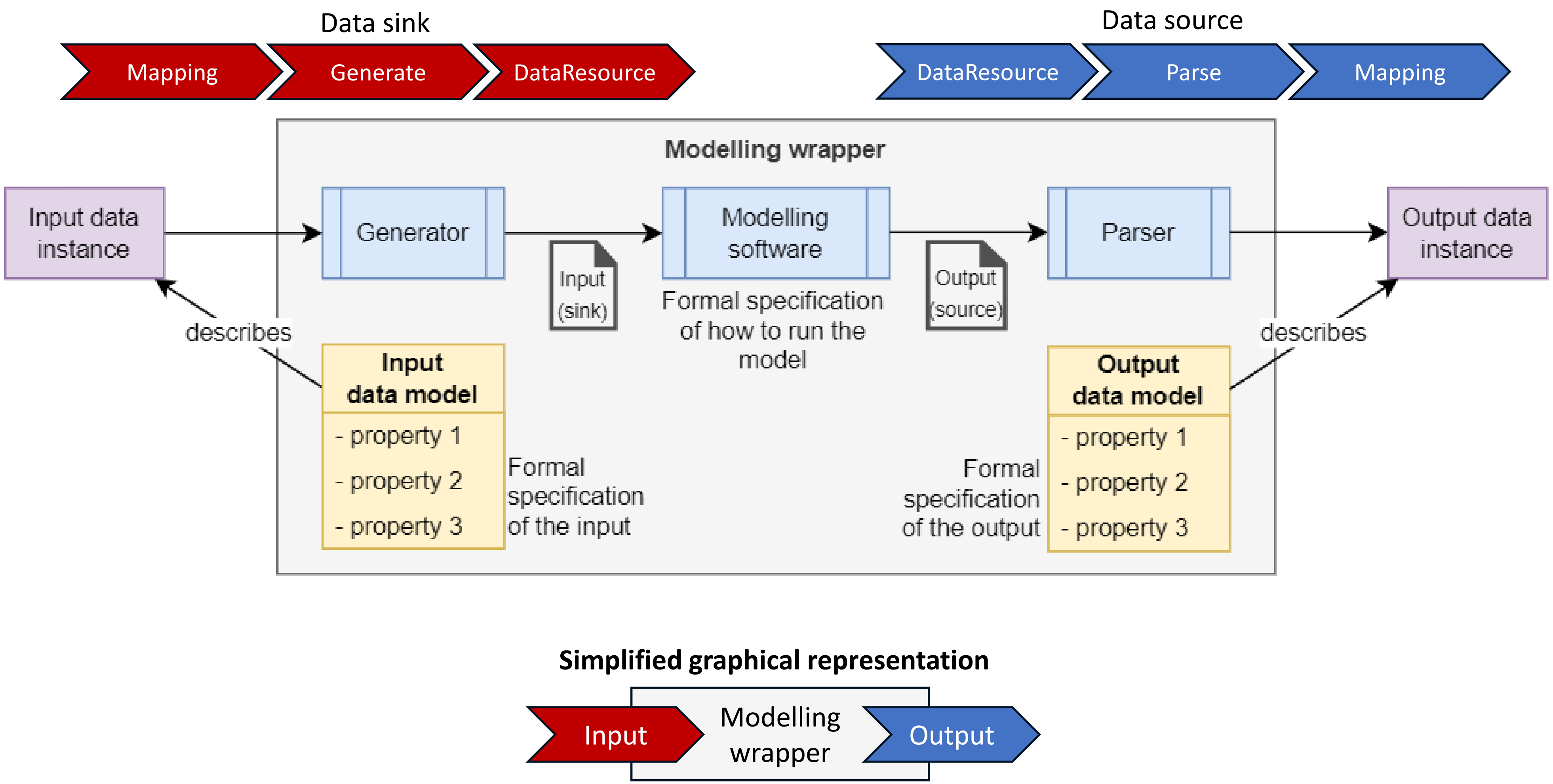
Summary

Levels of data documentation

… and what aspects of FAIR data they address



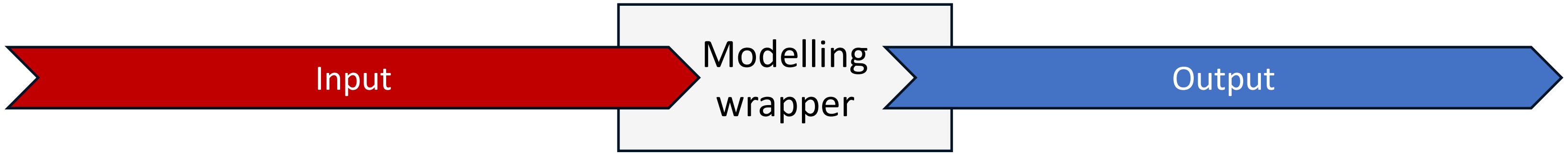
Documenting modelling wrappers



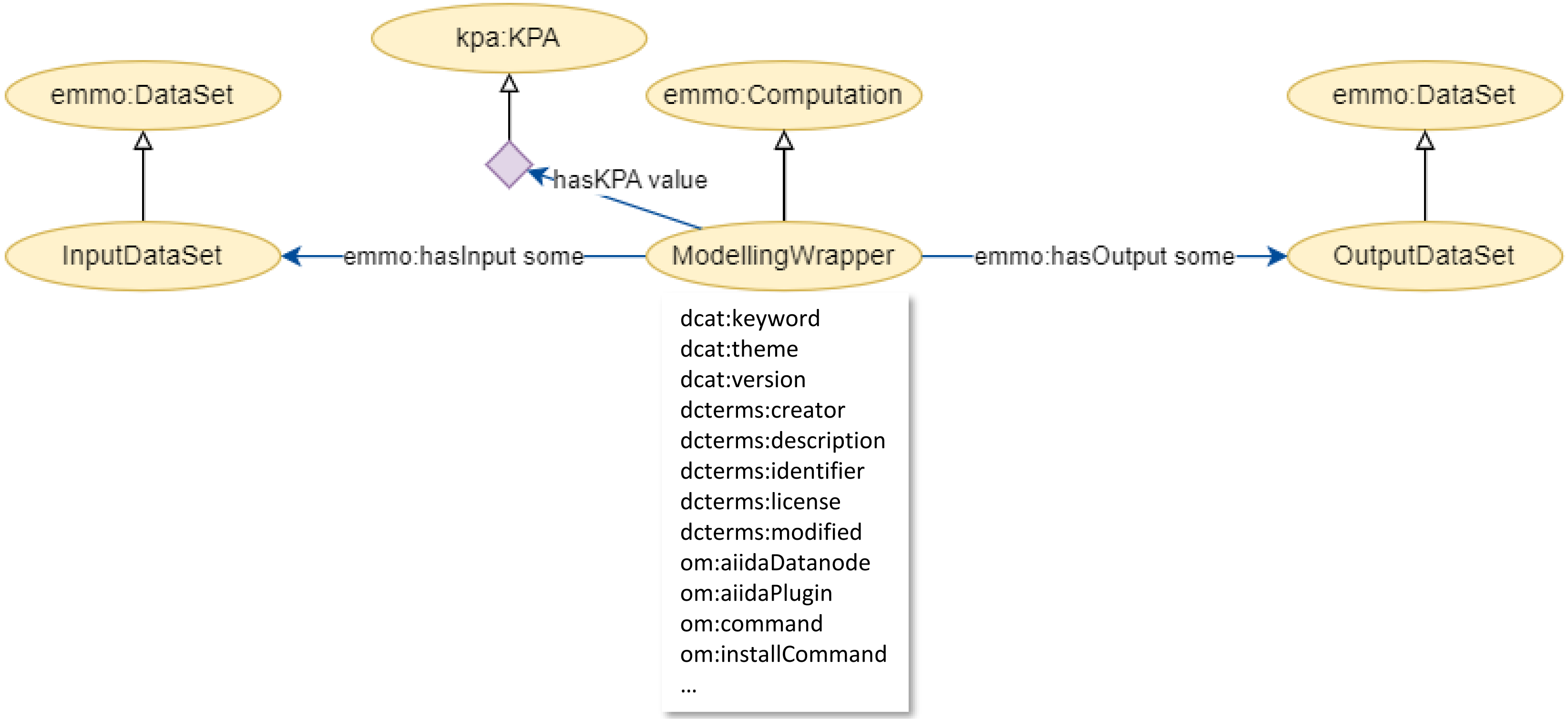
SINTEF

Representing a computation

Semantic documentation of output/input (interoperability)

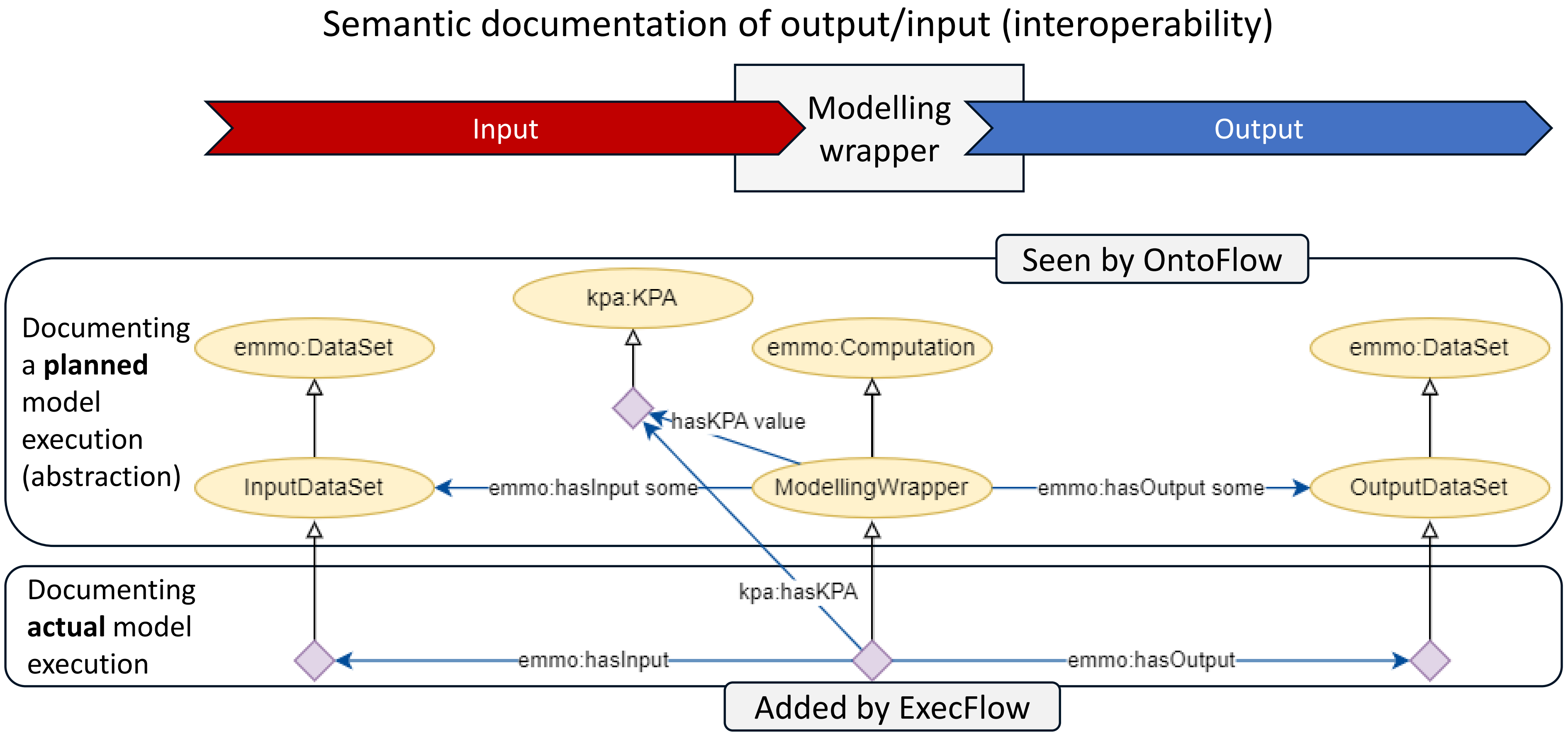


Documenting a **planned** model execution (abstraction)



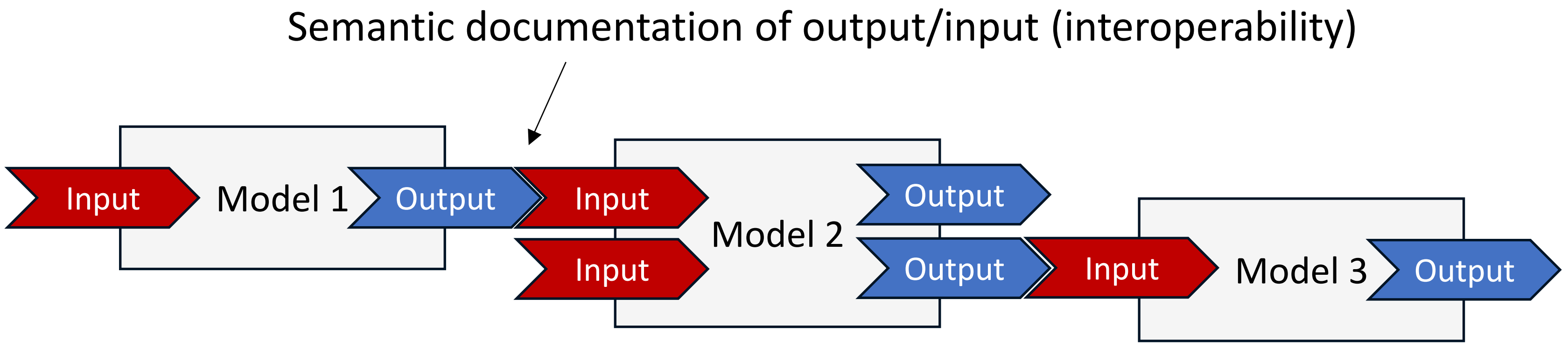
SINTEF

Representing a computation

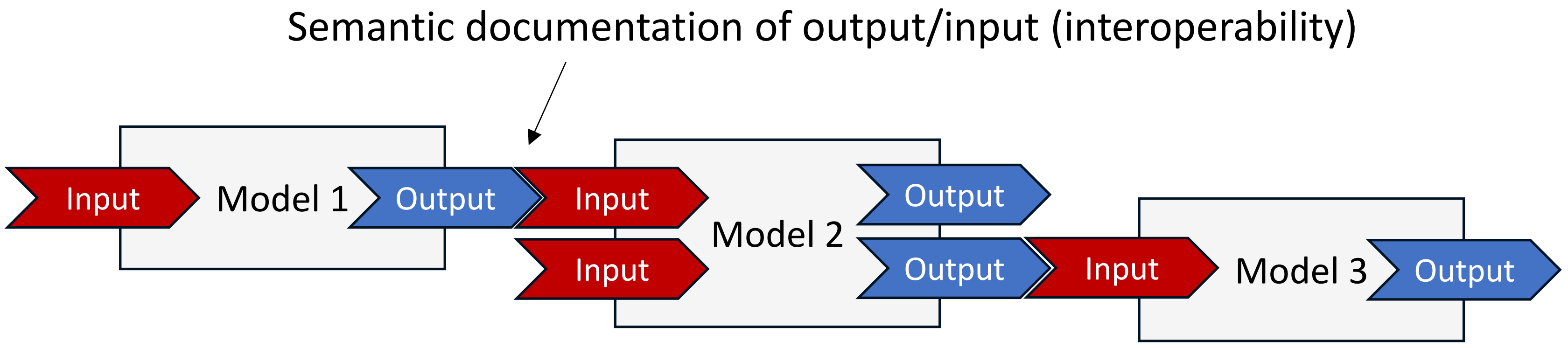


SINTEF

Representing a workflow



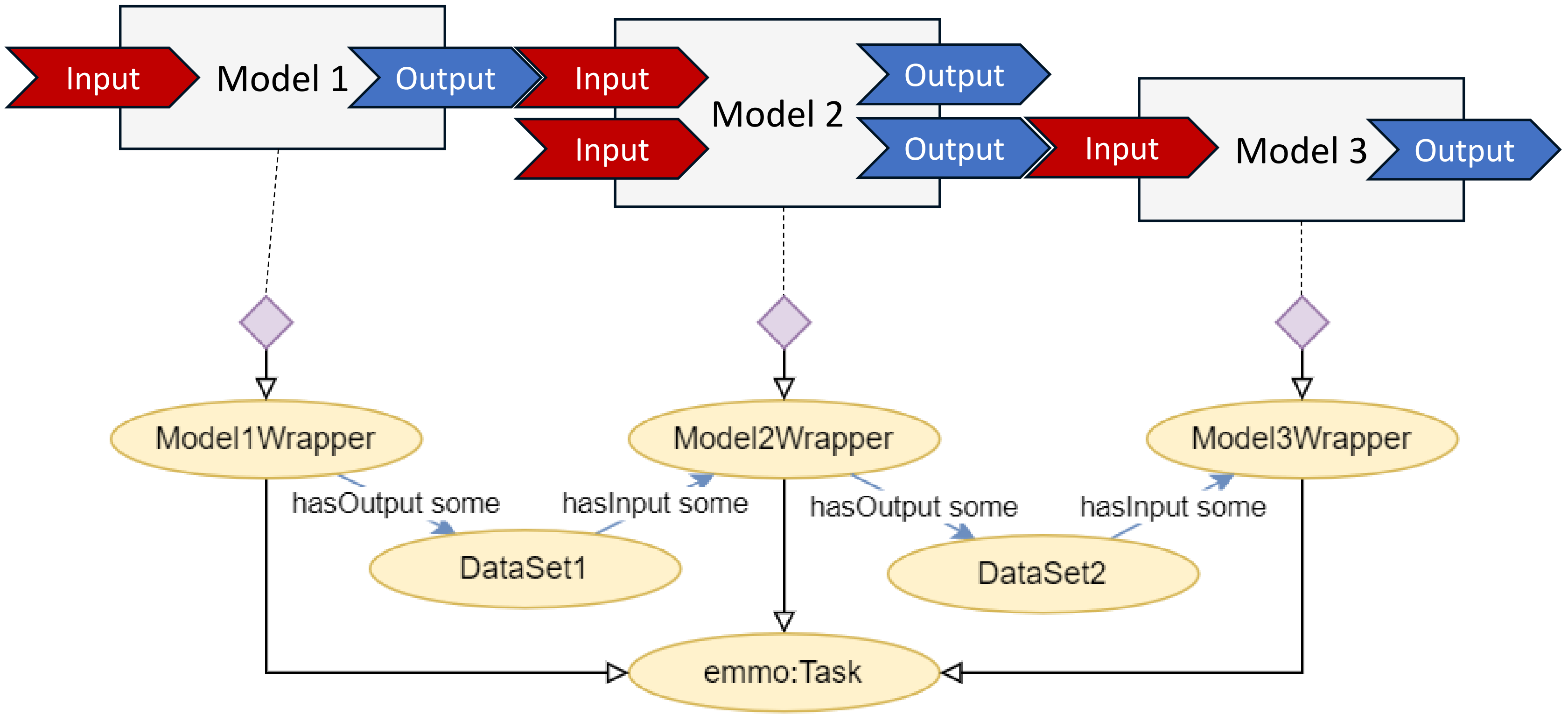
Representing a workflow



Build by OntoFlow

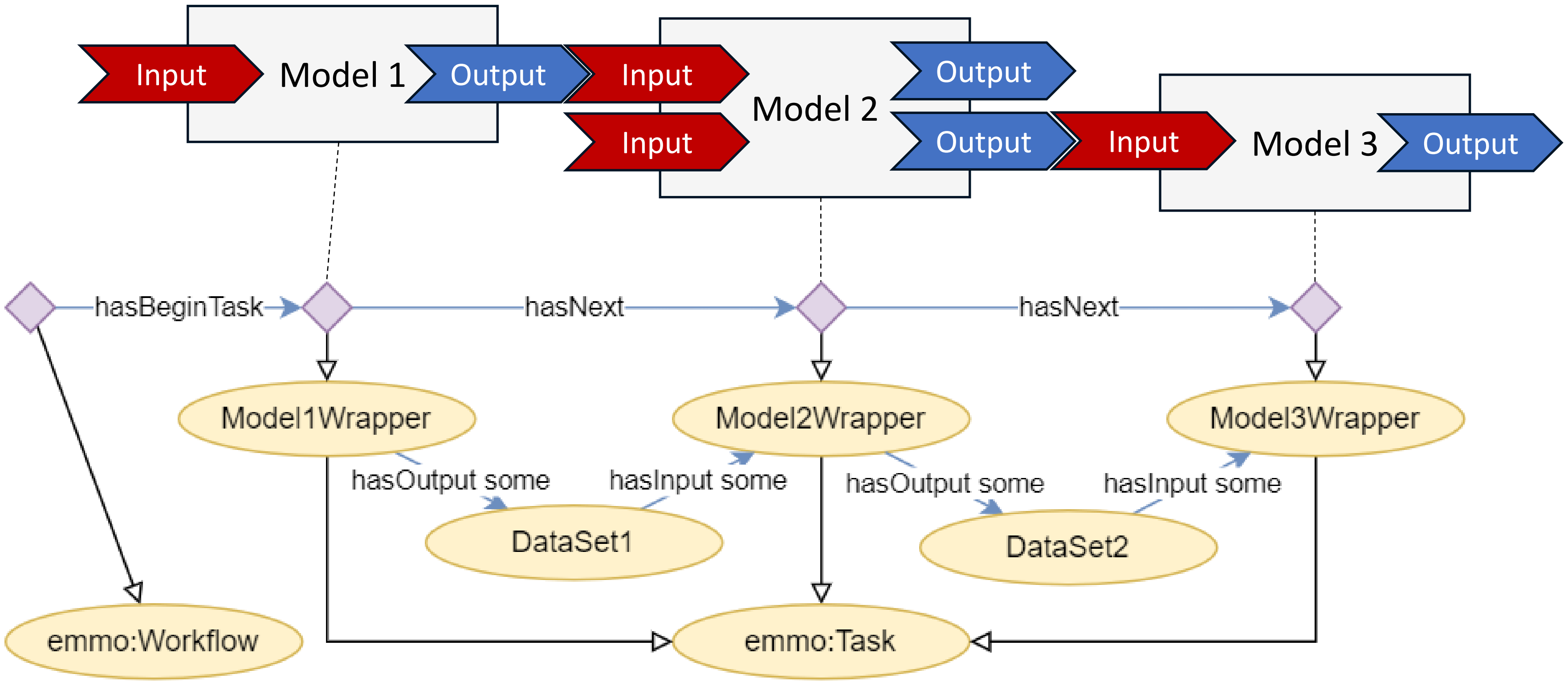
Executed by ExecFlow

Representing a workflow



SINTEF

Representing a workflow – next step



SINTEF

YAML documentation of a modelling wrapper

```yaml
version: 1

prefixes:
 ss3: "http://open-model.eu/ontologies/ss3#"

simulation_resources:

 ss3:AbaqusSimulation:
  # Name of the AiiDA plugin to use
  aiida_plugin: execflow.exec_wrapper

  # The command for running the simulation tool
  command: run_abaqus.sh

  # The commands for installing the wrapper to the simulation tool
  install_command: |
   pip install git+ssh://git@github.com:H2020-OpenModel/SS3_wrappers.git@master

  # Documentation of all Abaqus ouput:
  output: # Documentation of all Abaqus ouput as source partial pipeline in an oteapi pipeline
   # Documentation of Abaqus input for Aluminium material card
   ss3:DeformationHistory:
    - dataresource:
      mediaType: application/vnd.dlite-parse
      downloadUrl: abaqus_deformation_history.txt
      configuration:
        driver: plugin_abaqus_output
        datamodel: http://www.sintef.no/calm/0.1/AbaqusDeformationHistory

  # Documentation of all Abaqus input:
  input: # Documentation of all Abaqus input to be provided as sink pipeline by an oteapi pipeline
   # Documentation of Abaqus input for Aluminium material card
   ss3:AluminiumMaterialCard:
    - function:
      functionType: application/vnd.dlite-generate
      configuration:
       driver: plugin_abaqus_material
       # the name of the file is as expected by the Abaqus input template
       location: Section_materials_al.inp
       datamodel: http://www.sintef.no/calm/0.1/AluminiumMaterialCard

   # Documentation for Abaqus input for Concrete material card
```