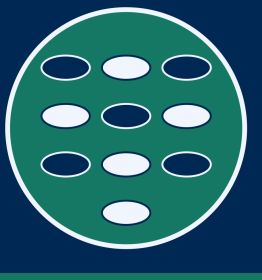
Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap



Digitalisering på Ås

DAT121
Introduction to data science

4 Good practice

4.1 Language of evidence
4.2 Interpretability and reproducibility
4.3 Data documentation

Schedule for DAT121 parts 4 and 5

Friday, 25th August 2023

10.15 – 11.00 lecture on good practice
11.15 – 12.00 **interest group sessions**
13.15 – 15.00 project work and tutorial

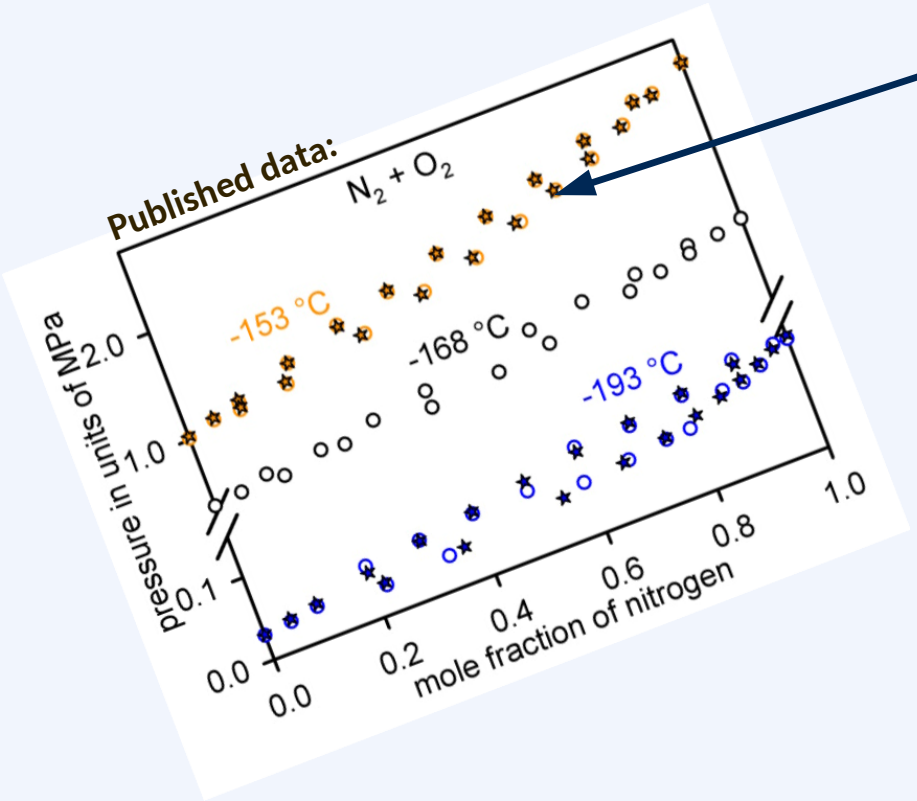
Monday, 28th August 2023

9.15 – 10.00 first multidimensionality lecture
10.15 – 10.?? Pangasia presentation in TF1-115
11.15 – 12.00 discussion and problem solving
13.15 – 15.00 project work and tutorial

Tuesday, 29th August 2023

9.15 – 10.00 Q&A session and discussion
10.15 – 11.00 second multidimensionality lecture
11.15 – 12.00 **interest group sessions**
13.15 – 15.00 project work and tutorial

Why do we need good practices?



What values did *x* and *p* have?

How was the data point obtained?

What is the margin of error, how was the error defined, and what software (or experimental setup) was used?

Good practice in managing research data:

Make all data **findable**, **accessible**, **interoperable**, and **reusable** (FAIR).

Why do we need good practices?

Norwegian University of Life Sciences

Recommendations

R1: definitions of concepts, metadata and data schemes
R2: creating semantic artefacts with open licenses
R3: associated documentation for semantic artifacts
R4: repositories of semantic artefacts
R5: minimum metadata model and cross walks discovery
R6: extensible options for disciplinary metadata
R7: apply a broad definition of data (datasets, workflows, lab protocols, software, methods, hardware design, etc.)
R8: clear protocols and building blocks for catalogues

Problems

Lack of (or overabundance of)

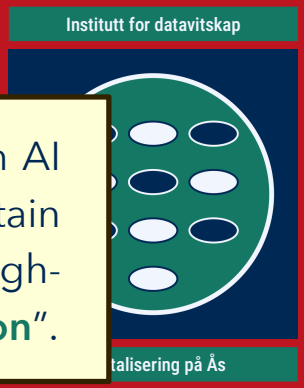
P1: explicit definitions
P2: common semantics (general ont
P3: reference repository
P4: common metadata scheme across communities
P5: metadata models

Needs

N1: principle approaches/tools for ontology and metadata schemes
N2: harmonisation across disciplines
N3: harmonisation of data of the same type
N4: federated access to existing research data repositories

EUROPEAN OPEN SCIENCE CLOUD

O. Corcho *et al.*, EOSC Interoperability Framework, doi:10.2777/620649, **2021**.



European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems. […] High-risk AI systems should therefore be accompanied by **relevant documentation**".

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

4 Good practice

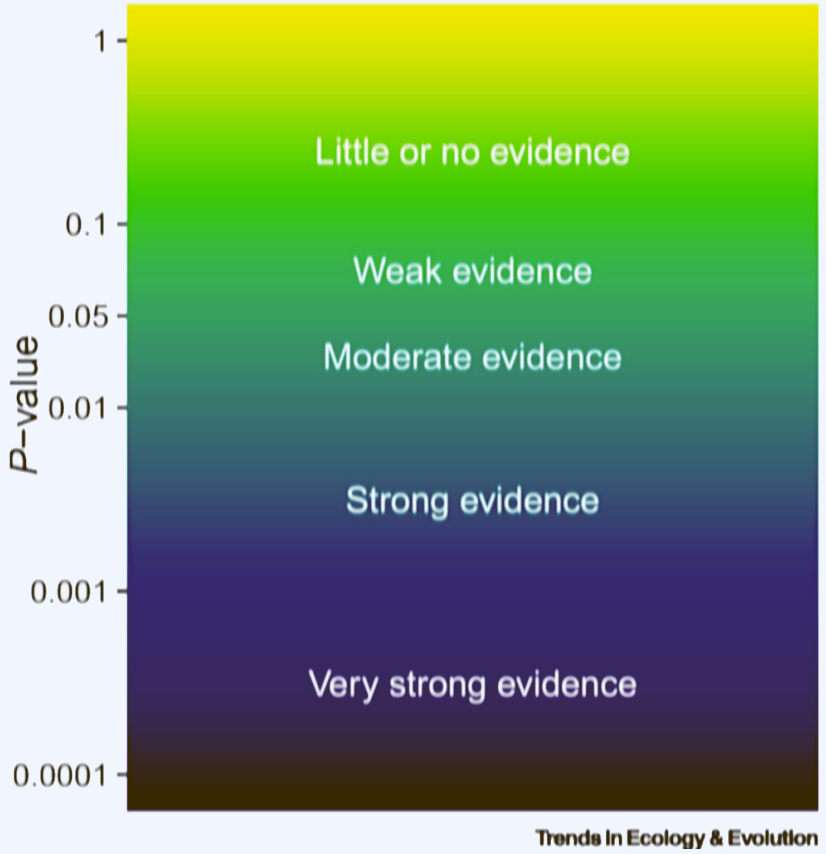
4.1 What is known from data?

Alternatives to the *p* value

There is always the risk of **statistical fallacies** when we overly rely on the *p* value.

"Eat elk meat to avoid heart attacks!"

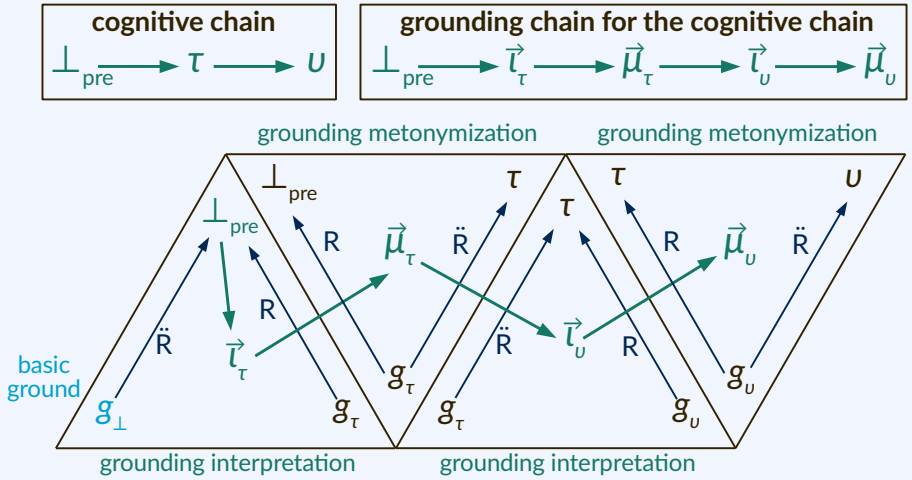
p-values → language of evidence[1]



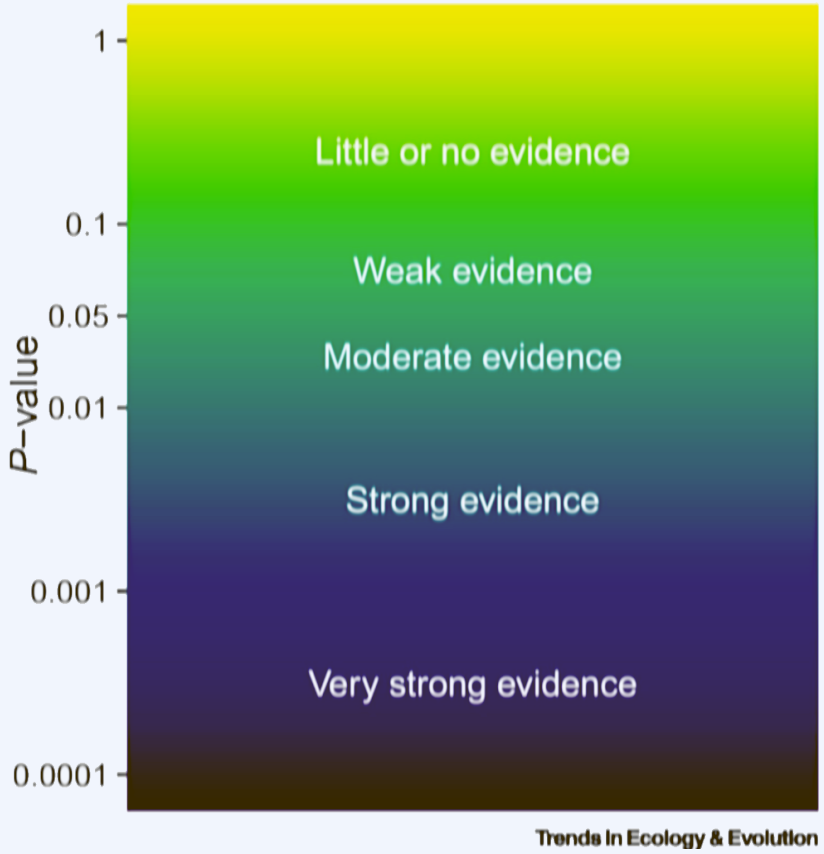
[1]S. Muff *et al.*, "Rewriting results sections in the language of evidence," doi:10.1016/j.tree.2021.10.009, **2022**.

Alternatives to the *p* value

The **epistemic grounding** of a research outcome is an explanation for why the scientific community *accepts that result as knowledge*;[2] or a rationale for why it *should be accepted* as knowledge.



p-values → language of evidence[1]



[1]S. Muff *et al.*, "Rewriting results sections in the language of evidence," doi:10.1016/j.tree.2021.10.009, **2022**.

[2]M. Horsch, B. Schembera, "Documentation of epistemic metadata […]", in *Proc. JOWO 2022* (CAOS), **2022**.

Epistemic opacity and metadata

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems.[1] […] High-risk AI systems should therefore be accompanied by **relevant documentation**".

[1]Systems with "high risk" include "safety components" related to "water, gas, heating, and electricity."

What are the epistemically relevant elements?

What is the relevant documentation that must accompany the AI systems?

Epistemic opacity and metadata

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems.[1] […] High-risk AI systems should therefore be accompanied by **relevant documentation**".

[1]Systems with "high risk" include "safety components" related to "water, gas, heating, and electricity."

Epistemic metadata:

a) "what **knowledge claim (KC)** φ has been formulated?,"
b) "where do the data and the claim come from?" (**provenance**),
c) "what **validity claim (VC)** was made about φ?,"
d) "why should we accept any of this?" (**grounding**).

Reproducibility, verification, and falsification

reproducibility

There are many definitions of reproducibility and replicability; see work by Hans Ekkehard Plesser (2018).

1) Reseacher *a* did *κ* and found *φ*.

2) Reseacher *b* did *γ*, which is very similar to *κ*, and found *ζ*, not very similar to *φ*.

3) Nobody disputes *a*'s integrity. Nobody disputes that *a* did *κ* and found *φ*.

Reproducibility claim (RC)

«Whenever the research process *κ*'' is carried out, it **must** lead to the outcome *φ*''.»

Reproducibility, verification, and falsification

Common formulation and schema for reproducibility claims (RCs):

«Whenever research process κ'' is carried out, it must lead to the outcome φ''.»

1) Reseacher *a* did κ and found φ.

   Here, *a* also made a **positive reproducibility claim ψ**.

2) Reseacher *b* did γ, **consistent with κ''**, and found ζ, **inconsistent with φ''**.

   Here, *b* made the **negative reproducibility claim $\neg\psi$**.

3) What is relevant there is the **contradiction between ψ and $\neg\psi$**.

provenance metadata κ
provenance paradata κ'

---

provenance orthodata $\kappa'' = \kappa - \kappa'$

«repeat κ, but no need to retain κ'»

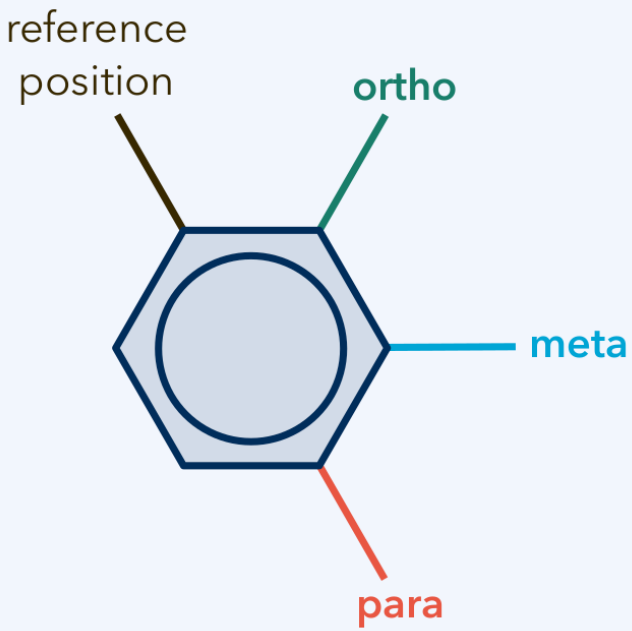
knowledge claim metadata φ
knowledge claim paradata φ'

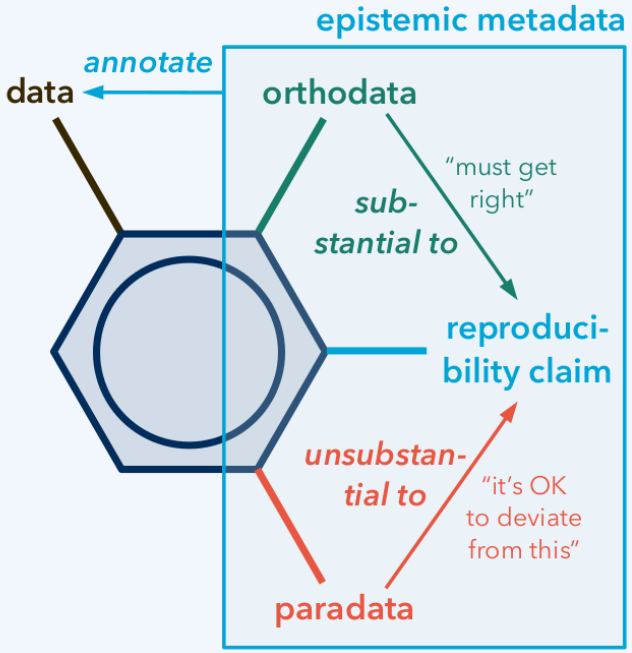
---

knowledge claim orthodata $\varphi'' = \varphi - \varphi'$

«obtain φ again, except for φ' maybe»

Reproducibility, verification, and falsification





provenance metadata κ
provenance paradata κ'

---

provenance orthodata $\kappa'' = \kappa - \kappa'$

«repeat κ, but no need to retain κ'»

knowledge claim metadata φ
knowledge claim paradata φ'

---

knowledge claim orthodata $\varphi'' = \varphi - \varphi'$

«obtain φ again, except for φ' maybe»

Norwegian Reproducibility Network (NORRN)

Our Mission

The Norwegian Reproducibility Network (NORRN) is a peer-led network that aims **to promote and enable rigorous, robust and transparent research practices in Norway**. We attempt to achieve this goal by establishing appropriate training activities, designing, and evaluating research improvement efforts, disseminating best practices, and working with stakeholders to ensure coordination of efforts across the sector. NORRN's activities span multiple levels, including researchers, librarians, institutions, and other stakeholders (e.g., funders and public authorities).

Researchers

We **support researchers** in educating themselves about open science practices, and founding local open science communities.

Initiatives

We **connect Reproducibility Initiatives** to a national network, and foster connections between them.

Institutions

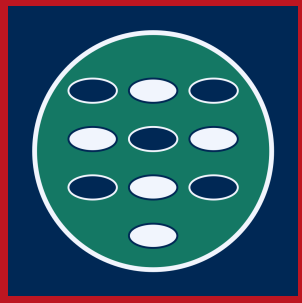
We **advise institutions** on how to embed open science practices in their work.

Stakeholders

We **represent the open science community** toward other stakeholders in the wider scientific landscape.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap



Digitalisering på Ås

4 Good practice

4.1 What is known from data?

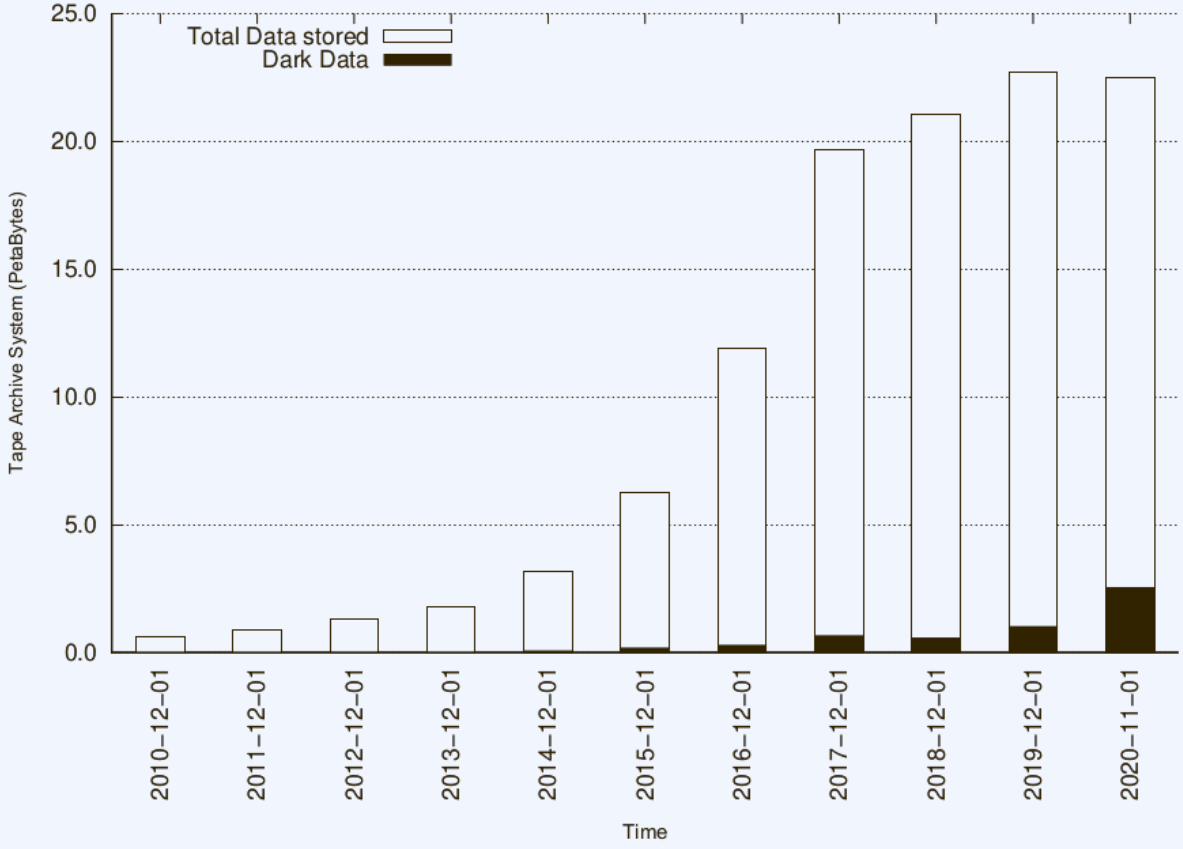
4.2 Data management principles

DAT121

25. august 2023

The challenge: Dark data

Dark data are data with an uncharacterized epistemic status.
In other words: We do not know what we know from and about the data.



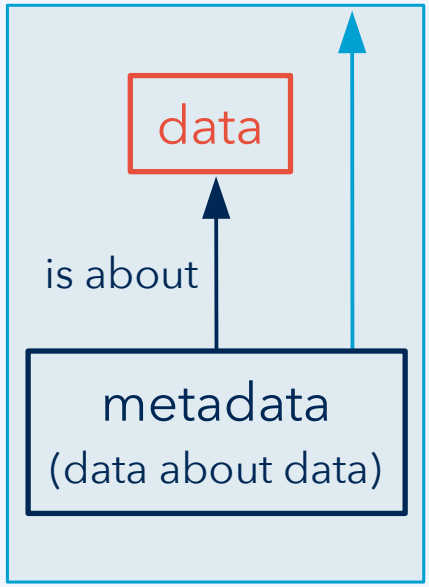
dark data

Flood of dark data:
More and more data are accumulated, but are dark – and useless.

Source: Björn Schembera, from work by Juan Durán and Björn Schembera.

Two traditions in data documentation

Challenge: Data and metadata need to become **explainable-AI-ready** (XAIR).



FAIR digital object[1, 2]

The librarian:

- Focus on **archival** and **curation**
- **Help humans** use digital artefacts
- Focus on **provenance**, like for artefacts in a museum, so humans understand where they come from

The engineer:

- Computers must **understand** what the digital artefacts mean
- **Focus on knowledge**
- FAIR digital objects[1, 2]
- Aim: Machine-actionability[2]

[1] I. Anders *et al.*, *FAIR Digital Object Technical Specification*, doi:10.5281/zenodo.7824713, **2023**.

[2] C. Weiland, S. Islam, *et al.*, *FDO Machine Actionability*, doi:10.5281/zenodo.7825649, **2023**.

FAIR principles[1] in detail

Norwegian University of Life Sciences

persistent identifier

Findability

F1. Globally unique **persistent identifiers (PID)**
F2. **Enriched with metadata**
F3. Data identifier included in metadata
F4. **Registered** in **searchable platform**

Accessibility

A1. **Retrievable from PID** via a standard protocol
A1.1. Open and freely implementable protocol
A1.2. … **authentication/authorization** if necessary
A2. Metadata remain accessible (beyond data)

Interoperability

I1. **Formal language** used for **knowledge representation**
I2. Metadata use **vocabularies** that are themselves FAIR
I3. Semantic web principles, **data can refer to other data**

Reusability

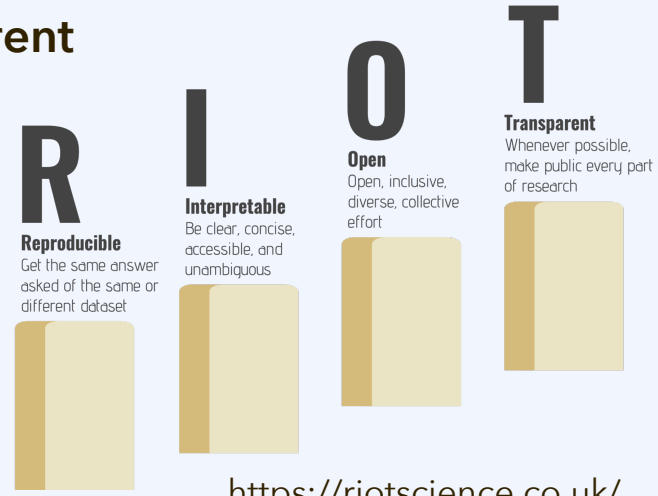
R1. Metadata include a plurality of accurate and relevant attributes
R1.1. Release data and metadata with an accessible **data usage license**
R1.2. Data are annotated with a detailed **provenance description**
R1.3. Relevant **disciplinary and community standards** are fulfilled

[1]M. D. Wilkinson *et al.*, "The FAIR Guiding Principles …," doi:10.1038/sdata.2016.18, **2016**.

RIOT principles

RIOT:[1] Reproducible, interpretable, open, transparent

- Origin: UK Reproducibility Network (UKRN)
- UKRN encouraged foundation of the other reproducibility networks, such as NORRN, the Norwegian Reproducibility Network
- Local "RIOT science clubs" were founded



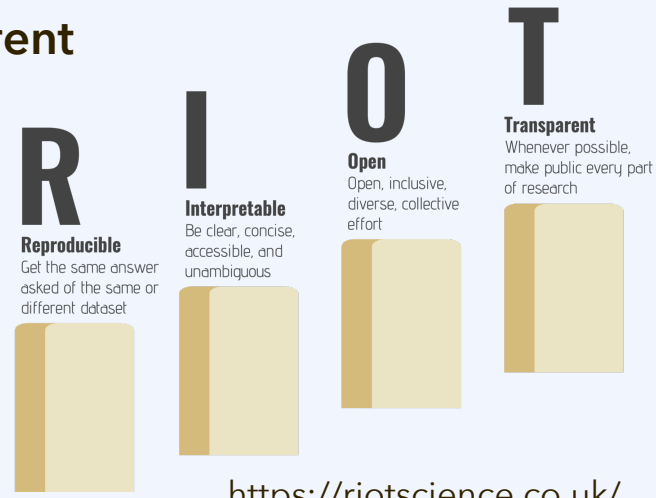
https://riotscience.co.uk/

[1]E. Ganley *et al.*, *BMC Res. Notes* **15**: 51, doi:10.1186/s13104-022-05932-5, **2022**.

RIOT, FAIR, and CARE principles

RIOT:[1] Reproducible, interpretable, open, transparent

- Origin: UK Reproducibility Network (UKRN)
- UKRN encouraged foundation of the other reproducibility networks, such as NORRN, the Norwegian Reproducibility Network
- Local "RIOT science clubs" were founded



https://riotscience.co.uk/

CARE:[2] Collective benefit, authority to control, responsibility, ethics

- Origin: Global Indigenous Data Alliance
- Uptake supported by the Research Data Alliance
- Orientation: Sovereignty and epistemic justice

https://www.gida-global.org/care/

[1]E. Ganley *et al.*, *BMC Res. Notes* **15**: 51, doi:10.1186/s13104-022-05932-5, **2022**.
[2]S. Russo Carroll *et al.*, *Sci. Data* **8**: 108, doi:10.1038/s41597-021-00892-0, **2021**.

FAIR ontologies

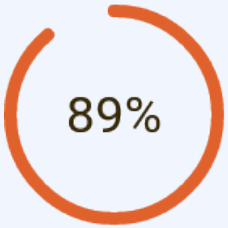
In dealing with data, we should make all our content FAIR. Leading the way, first and foremost the *ontologies* themselves *must also be FAIR*.

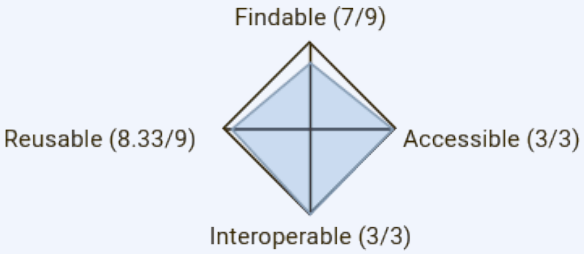
In an exercise from 2021/22, over 50 ontologies from industrially relevant domains were checked against minimum standards for FAIRness.

How many do you think were successful at fulfilling the minimum standard?

The **Foops! validator** checks ontologies for FAIRness. It also helps developers make their ontologies FAIR by providing constructive feedback.

example feedback from Foops![1]



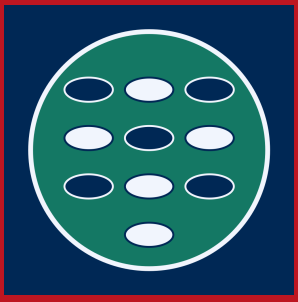


URL:
https://foops.linkeddata.es/FAIR_validator.html

[1]D. Garijo *et al.*, *Proc. ISWC 2021 Posters/Demos/Industry*, p. 321, **2021**.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap



Digitalisering på Ås

4 Good practice

4.1 What is known from data?

4.2 Data management principles

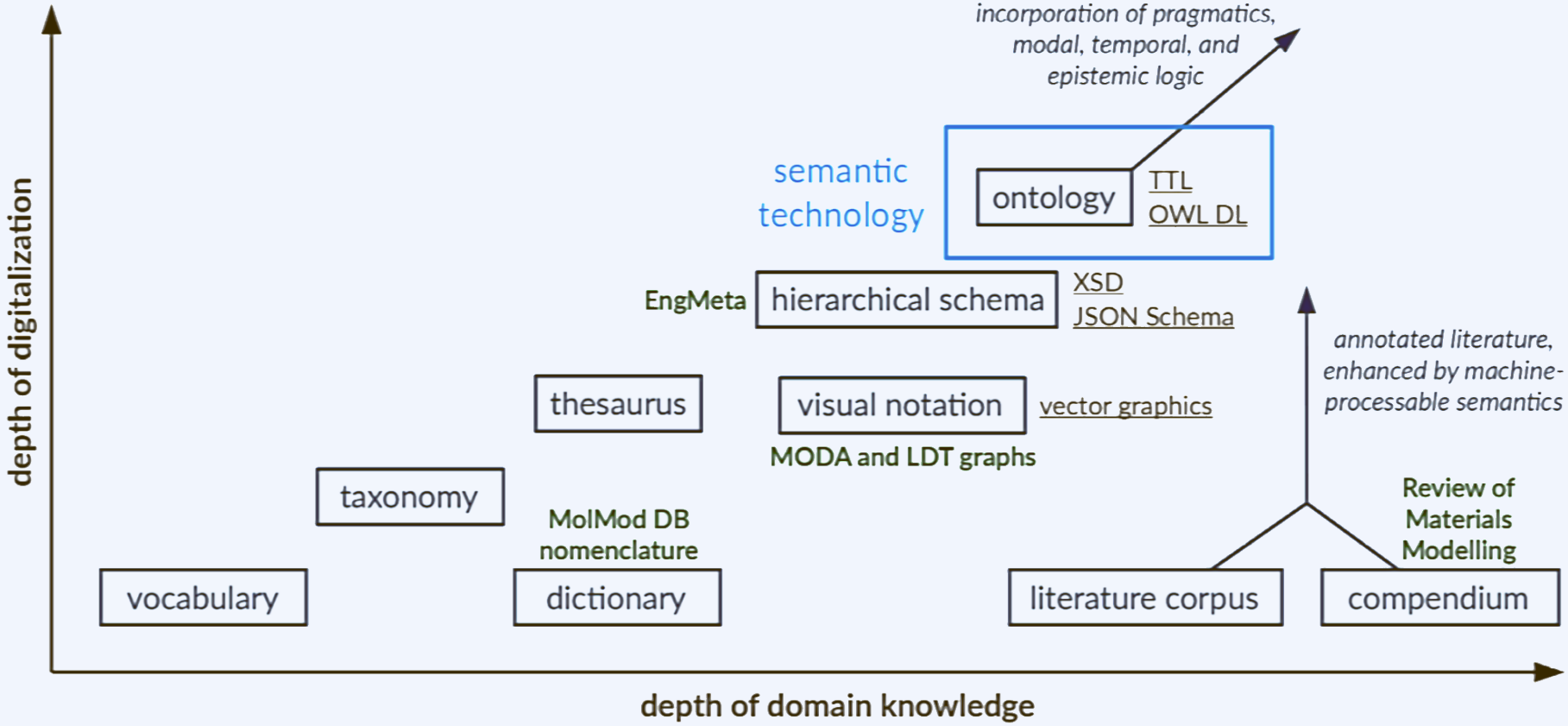
4.3 Ontology engineering practice

DAT121

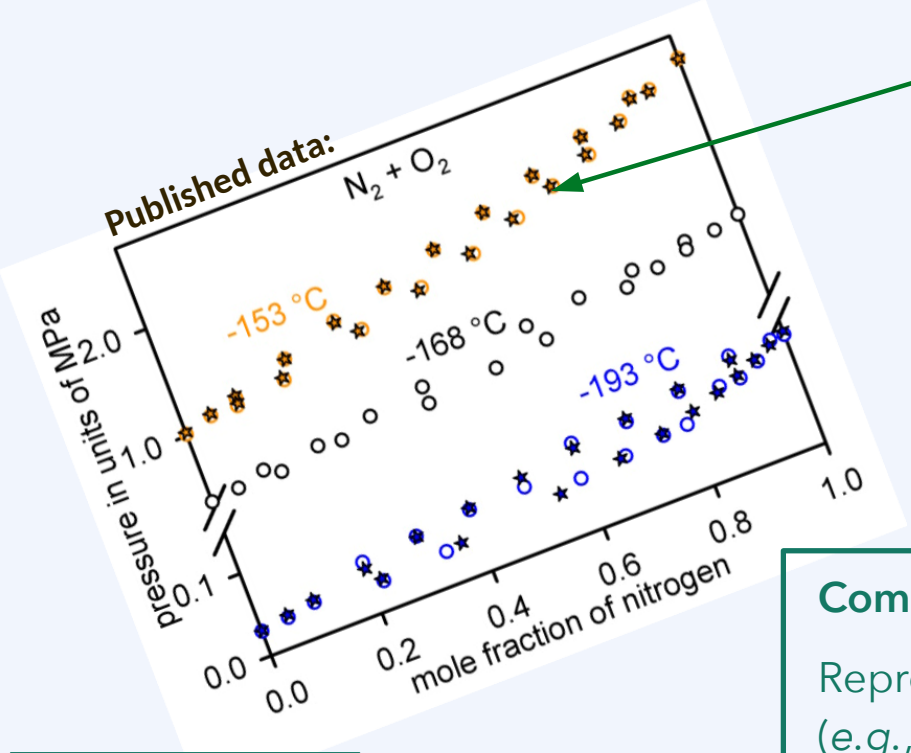
25. august 2023

Agreed metadata through standardization

Types of **semantic artefacts**, also referred to as **metadata standards**:



Bottom-up approach: Competency questions



What values did *x* and *p* have?

How was the data point obtained?

What is the margin of error, how was the error defined, and what software (or experimental setup) was used?

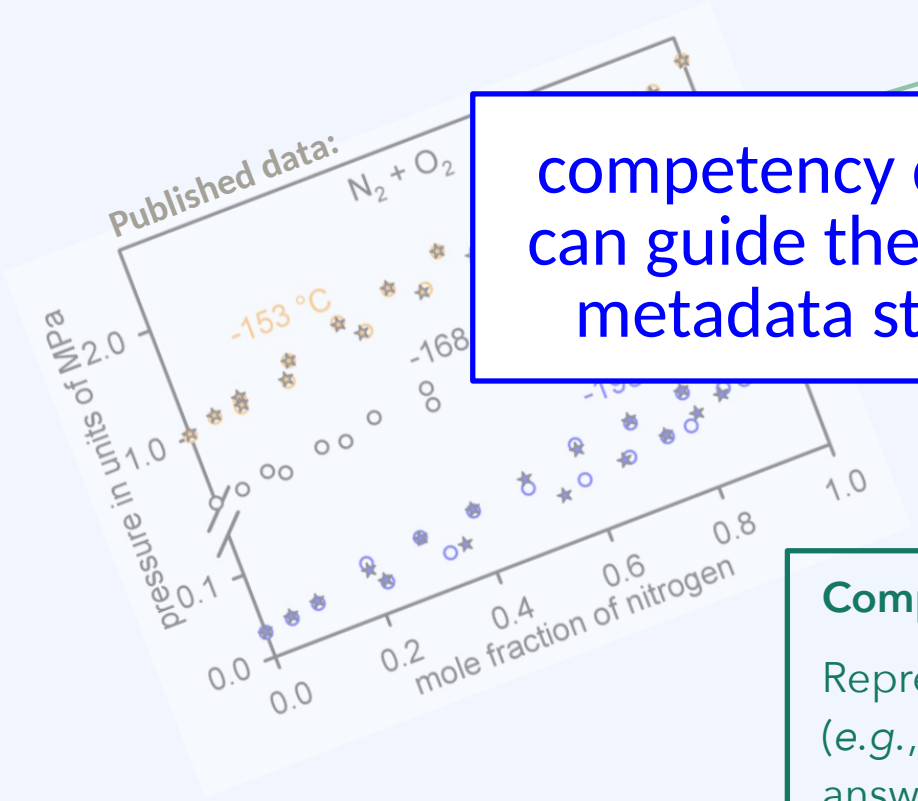
Competency questions:[1, 2]

Representative queries about data (*e.g.*, for metadata), to be competently answered by a knowledge base.

competency question

[1]M. Grüninger, M. S. Fox, in *Benchmarking: Theory and Practice*, doi:10.1007/978-0-387-34847-6_3, **1995**.
[2]C. Bezerra *et al.*, *Learning Nonlin. Models* **12**(2): 115-129, doi:10.21528/lnlm-vol12-no2-art4, **2014**.

Bottom-up approach: Competency questions



competency questions can guide the design of metadata standards

What values did *x* and *p* have?

data point obtained?

argin of error, how was
ned, and what software
(or experimental setup) was used?

Competency questions:[1, 2]

Representative queries about data (*e.g.*, for metadata), to be competently answered by a knowledge base.

[1]M. Grüninger, M. S. Fox, in *Benchmarking: Theory and Practice*, doi:10.1007/978-0-387-34847-6_3, **1995**.
[2]C. Bezerra *et al.*, *Learning Nonlin. Models* **12**(2): 115-129, doi:10.21528/lnlm-vol12-no2-art4, **2014**.

Top-down approach: Foundational ontology

foundational ontology

A foundational ontology provides a general structure to the semantics of any kind of potential information content. (Or at least it claims to.)

Benefits for users:

- You don't have to redevelop the most abstract concepts. It was already done by the foundational ontology, thoroughly analysed and tested.
- Other ontology developers will already know these high-level concepts.
- You can more easily align (i.e., match and connect) your ontology to other developers' ontologies, if they use the same foundational ontology.

DOLCE

Descriptive Ontology for Linguistic and Cognitive Engineering

http://www.loa.istc.cnr.it/dolce/overview.html

EMMO

Elementary Multiperspective Material Ontology

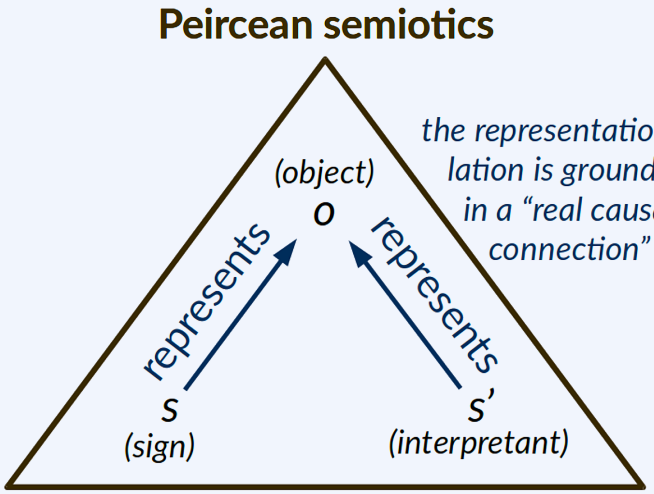
https://emmo-repo.github.io/

BFO

Basic Formal Ontology

https://basic-formal-ontology.org/

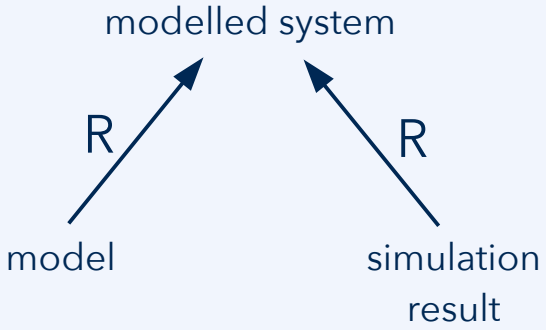
Foundational ontology: EMMO



C. S. Peirce

Elementary Multiperspective Material Ontology[1, 2]

1) **Taxonomy:**
   Conceptual hierarchy (subclass relation)
2) **Mereocausality:**
   Spatiotemporal parthood and connectivity
3) **Semiotics:**
   Representation of physical entities by signs

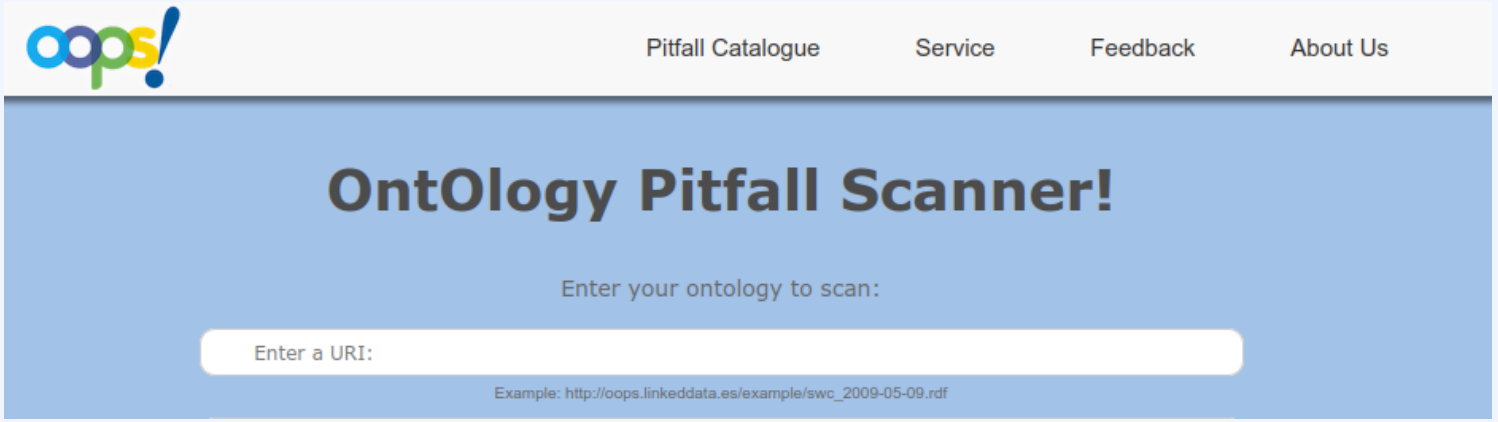


"represents" or **"is sign for"** is here abbreviated by R

[1]H. A. Preisig *et al.*, doi:10.23967/wccm-eccomas.2020.262, no. 262 in *Proc. ECCOMAS 2020*, **2021**.
[2]S. Clark *et al.*, *Adv. Energ. Mat.* 12(17), 2102702, doi:10.1002/aenm.202102702, **2022**.

Ontology design pitfalls

The **Oops! Ontology Pitfall Scanner** helps ontology desingers avoid technical shortcomings and mistakes. URL: https://oops.linkeddata.es/



- **Critical** : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** : It is not really a problem, but by correcting it we will make the ontology nicer.

[Expand All] | [Collapse All]

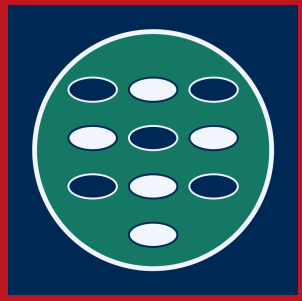
Results for P11: Missing domain or range in properties.	**195 cases | Important**
Results for P13: Inverse relationships not explicitly declared.	**45 cases | Minor**
Results for P36: URI contains file extension.	**ontology* | Minor**
SUGGESTION: symmetric or transitive object properties.	**5 cases**

example feedback from Oops![1]

[1]M. Poveda *et al.*, *Int. J. Semant. Web Inform. Sys.* **10**: 7, doi:10.4018/ijswis.2014040102, **2014**.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap



Digitalisering på Ås

Conclusion

DAT121

25. august 2023

Norwegian University of Life Sciences

Related research activities

Epistemic Metadata for Computational Engineering Information Systems

Martin Thomas HORSCH [a,b,1] Silvia CHIACCHIERA [b]
Gabriela GUEVARA CARRIÓN [c] Maximilian KOHNS [d] Erich A. MÜLLER [e]
Denis ŠARIĆ [c] Simon STEPHAN [d] Ilian T. TODOROV [b] Jadran VRABEC [c]
Björn SCHEMBERA [f]

[a] *Norwegian University of Life Sciences, Faculty of Science and Technology, Department of Data Science, Drøbakveien 31, 1430 Ås, Norway*
[b] *UK Research and Innovation, STFC Daresbury Laboratory, Scientific Computing Department, Keckwick Ln, Daresbury WA4 4AD, UK*
[c] *Technische Universität Berlin, Thermodynamics, Ernst-Reuter-Platz 1, 10587 Berlin, Germany*
[d] *Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Laboratory of Engineering Thermodynamics, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany*
[e] *Imperial College London, Department of Chemical Engineering, South Kensington Campus, London SW7 2AZ, UK*
[f] *University of Stuttgart, Institute of Applied Analysis and Numerical Simulation, Pfaffenwaldring 57, 70569 Stuttgart, Germany*

Abstract. Digitalization is the main priority for innovation in the engineering sciences at present. This includes making the knowledge from scientific research data machine-actionable so that it can be integrated and analysed with minimal human intervention. Computational engineering has been advancing on this path for some time; *e.g.*, FAIR digital objects are gaining momentum as a paradigm for communicating data and metadata. Despite this, the depth of digitalization often remains too shallow, with annotations that are only of use to a human reader. In addition, digital infrastructures and their metadata standards are tedious to use: They require too much effort from researchers; *e.g.*, for providing input that contributes nothing to an automated reuse of knowledge. These two shortcomings, lack in depth and excess in breadth, are related. Addressing these gaps, the present contribution discusses metadata standardization efforts targeted at documenting the knowledge status of data; the required annotation is referred to as epistemic metadata. It is discussed how a metadata schema for knowledge and reproducibility can be designed such as to be user-friendly and flexible enough to apply to a spectrum of circumstances and types of replicability and consistency checks. These developments are positioned in the context of a recent case study on a sample of journal articles and knowledge claims from the domain of molecular modelling and simulation.

Keywords. Applied ontology, epistemic metadata, process data technology.

INPRODAT
https://www.inprodat.de/

EMMC
https://emmc.eu/

ONTO COMMONS
ONTOLOGY-DRIVEN DATA DOCUMENTATION FOR INDUSTRY COMMONS
https://ontocommons.eu/

DOME 4.0
https://dome40.eu/

Glossary terms

Proposed glossary[1] terms:

- How do we best define them? Is the definition controversial?
- What is the best translation into Norwegian bokmål/nynorsk?
- Are there more key concepts that would require an agreed definition?

persistent identifier

dark data

competency question

foundational ontology

(also: top-level ontology)

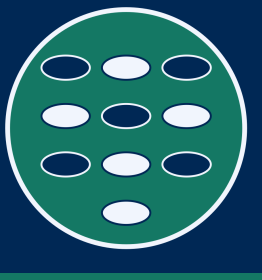
How much or what kind of reproducibility do you need to reuse previous work?

reproducibility

[1]https://home.bawue.de/~horsch/teaching/dat121/glossary-en.html

Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap



Digitalisering på Ås

DAT121
Introduction to data science

4 Good practice

4.1 Language of evidence
4.2 Interpretability and reproducibility
4.3 Data documentation