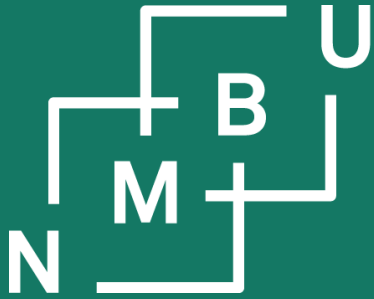# DAT390

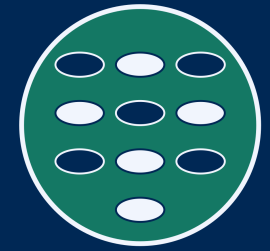# Data science seminar

3   Research methodology

3.1   Going beyond the state of the art
3.2   Research data management

3   Methodology

3.1   Beyond state of the art

23rd September 2024

How to be "beyond-the-art by design"

Master's Thesis 2024   30 ECTS
Faculty of Science and Technology

Assessing the Impact of Price Signals on Electricity Consumption Flexibility in Households: Evidence from Oslo, Norway

Hedda Kleven Berg
Environmental Physics and Renewable Energy
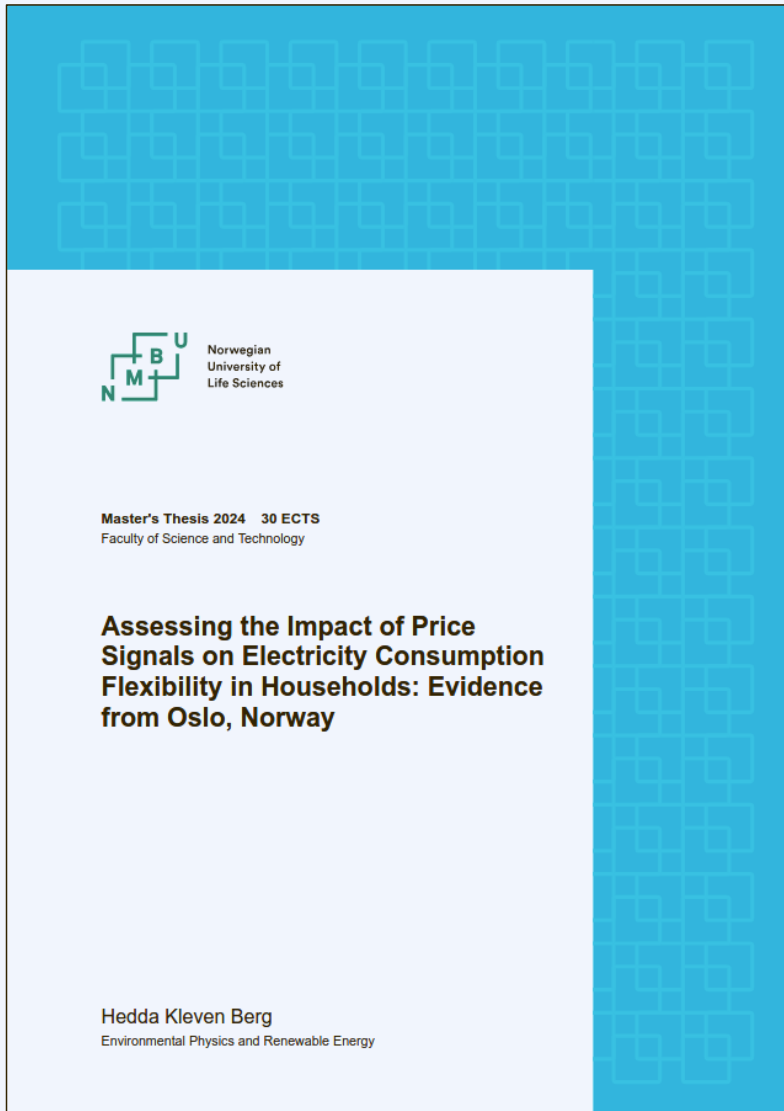
Be specific, work on a narrow topic.

It is easy to confirm that nobody else ever worked on this exact topic.

Also good: It is best if already from the title itself it becomes clear in what way the work is novel.

However: Being specific is not enough, we must **learn something novel** from it.

This can be something generalizable, but here it could also mean learning what is special about people in Oslo.

3

How to be "beyond-the-art by design"
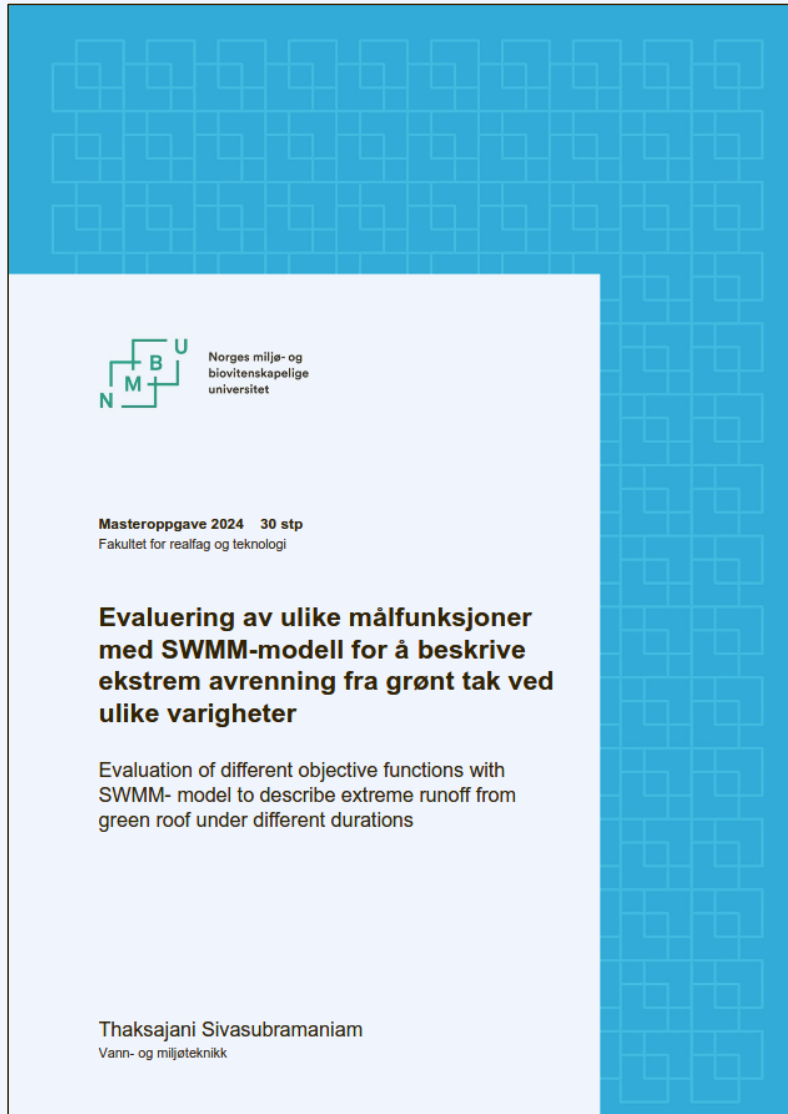
Use recently developed methods

Millions of people have done scientific research, how do you know yours is something nobody else did?

The SWMM model was developed in 2015, so it was possible for the author to evaluate the complete body of literature on it.

Your advisor should point you to recent methods or new approaches where it is very easy to do novel work that advances our body of knowledge.
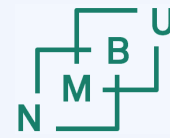
About the 2nd "nearly finished" report

From the "nearly finished report" on (deadline 8th November), the work is expected to <u>cover all aspects</u>, not just reviewing the state of the art.

It must in this sense be <u>complete</u>.

What distinguishes a "nearly finished" DAT390 report from the master thesis?
- The **master thesis** needs to report on the finalized research <u>and results</u>.

- The **master thesis** is a comparably long document.

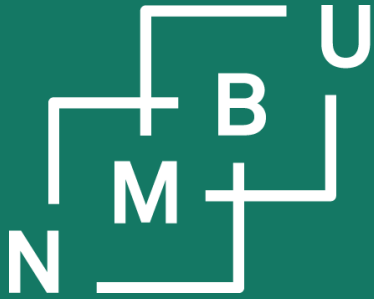# About the 2nd "nearly finished" report

See the UHR's recommended **standardized assessment form** for master theses.

What distinguishes a "nearly finished" DAT390 report from the master thesis?

– The **master thesis** needs to report on the finalized research <u>and results</u>. The **DAT390 report** is not expected to contain any final results. However, **preliminary results** or a **feasibility study** are needed.

– The **master thesis** is a comparably long document. The **DAT390 report** is like a conference paper, up to 12 pages (+ literature), 11pt, A4 paper.

– There is only one difference between the final **DAT390 report** and the **"nearly finished" report:** The DAT390 report determines your character grade in DAT390. The "nearly finished" report is not graded.

What are the **criteria for evaluating a master thesis**? See UHR document.

3 Methodology

23rd September 2024

Institutt for datavitskap
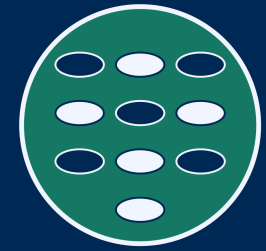
Digitalisering på Ås

Noregs miljø- og biovitskaplege universitet

NMBU's guidelines for research data management

Research data management guidelines document

(Approved by NMBU's rector in 2018)

- Principle "**open as standard**"

- Comply with international standards for research data management

- Comply with the **FAIR principles**.

- **Research data/metadata** should be available, searchable, and reusable.

- Furthermore, "interoperable" means that both data and metadata must be **manageable for machines** and that a **consistent vocabulary** is used.

NMBU's guidelines for research data management

Research data management guidelines document

(Approved by NMBU's rector in 2018)

    1) Research data must be **stored safely**.

    2) Research data should not (only) be stored on a local storage.

    3) Follow agreed **good practices**, regulations, guidelines, and the law.

    4) Persistent long-term **data preservation** for a minimum of ten years.

    5) The research data must be **annotated with metadata**.

    6) A **data management plan** (DMP), describing how data will be managed.

    7) Open and available (**open-access**) research data.

    8) Protect **personal data**.
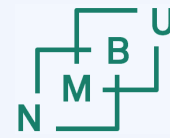
Compliance with GDPR

Concerning personal data, we need to comply with GDPR, and therefore:

– You need to make sure that there is a line of responsibility connecting your work to the **Data Protection Officer** (DPO), also *personvern-ombod*, of the organization. Hanne P. Gulbrandsen is NMBU's DPO.

GDPR: General data protection regulation. (Link to PDF document.)
This is implemented in Norway through personopplysningsloven (see document).

SIKT – Kunnskapssektorens tjenesteleverandør must be notified about use of personal data in your research project.

[1]https://gdpr.eu/data-protection-impact-assessment-template/?cn-reloaded=1

Data protection impact assessment

Concerning personal data, we need to comply with GDPR, and therefore:

- You need to make sure that there is a line of responsibility connecting your work to the **Data Protection Officer** (DPO), also *personvern-ombod*, of the organization. Hanne P. Gulbrandsen is NMBU's DPO.

- Your work may require a **Data Protection Impact Assessment** (DPIA) …
  - «if you're using new technologies»,
  - «data […] used to make automated decisions about people»,
  - «if you're tracking people's location or behavior», «monitoring a publicly accessible place» or «processing children's data», *etc.*

- You need **freely given**, **specific**, **informed**, and **unambiguous consent**.
  - Be aware of simultaneous requirements from NMBU's RDM policy.

SIKT – Kunnskapssektorens tjenesteleverandør must be notified about use of personal data in your research project.

[1]https://gdpr.eu/data-protection-impact-assessment-template/?cn-reloaded=1

Data protection impact assessment

Criteria from NTNU's knowledge base:[1]

1) "Evaluation or assigning a point score, included profiling and prediction […]
2) Automated decisions with legal or similarly significant consequences […]
3) Systematic monitoring/observation
4) Special categories of personal data [… *e.g.*] of a highly personal nature […]
5) Personal data are processed at a large scale […]
   > *«It is unclear what constitutes large scale.»*
6) Matching or integration of datasets […]
7) Personal data on vulnerable subjects […]
8) Innovative use or application of a new technical or organizational approach […]
   > *«in relation to the previously reached level of technical knowledge»*
9) Whenever the processing *«deprives those registered of the possibility to exercise a right or make use of a service or an agreement.»* This includes processing that aims at allowing, changing, or denying the registered person access to a service or enter an agreement."

[1]https://i.ntnu.no/wiki/-/wiki/Norsk/Vurdere+personvernkonsekvenser

Data protection impact assessment

Checklist from GDPR.EU

DPIA template from NTNU

Tools for drafting a data management plan (DMP)

Institutional templates for DMPs
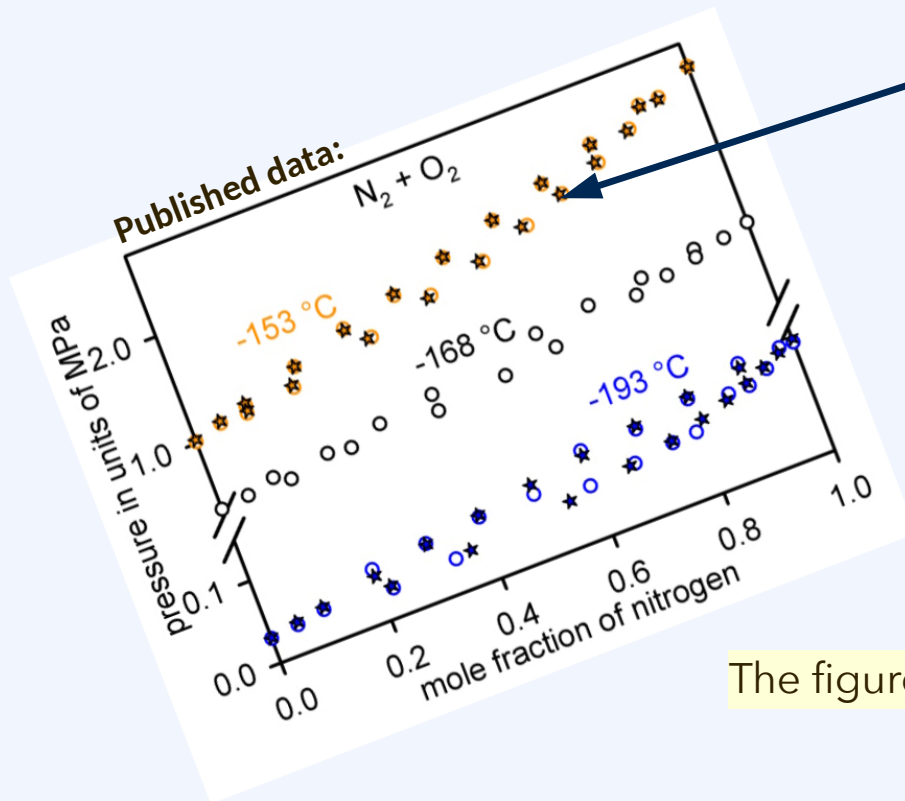
Example: DMP template for Horizon Europe projects

Data Stewardship Wizard (DSW)

Why do we need good practices?

Published data:

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?

The figure is not enough – **metadata** are needed!

Good practice in managing research data:

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

FAIR principles[1-3]

Summary from the guidance document by Direktoratet for e-helse:[3]

Findability: «That there is information in the form of **metadata**, making it simple for both humans and machines to find **datasets**.»

Accessibility: «That the **data and metadata** are arranged in a way that makes it simple for humans and machines **to obtain datasets**.»

Interoperability: «That **metadata and variables** are described in accordance with international **standards**, terminologies, and classifications.»

Reusability: «That there is information in the form of **metadata** that describe the **conditions for reuse** of the dataset, how the dataset has been edited, and the quality of the data.»

[1]M. D. Wilkinson *et al.*, "The FAIR Guiding Principles …," doi:10.1038/sdata.2016.18, **2016**.
[2]https://www.uio.no/for-ansatte/arbeidsstotte/forskningsstotte/forskningsdata/tema/fair.html (norsk versjon, UiO).
[3]Direktoratet for e-helse, Veileder for bruk av FAIR-prinsippene for helsedatakilder, **2020**.

FAIR principles in detail[1-3]

persistent identifier

Findability (gjenfinnbarheit)

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

Accessibility (tilgjengelegheit)

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 (samhandlingsevne)

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**

metadata

Reusability (gjenbruksevne)

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**.
[2]https://www.uio.no/for-ansatte/arbeidsstotte/forskningsstotte/forskningsdata/tema/fair.html (norsk versjon, UiO).
[3]Direktoratet for e-helse, Veileder for bruk av FAIR-prinsippene for helsedatakilder, **2020**.

What are metadata?

<u>Dataverse.NO</u>

18

Agreed metadata by standardization

Types of **semantic artefacts**, also referred to as **metadata standards**:
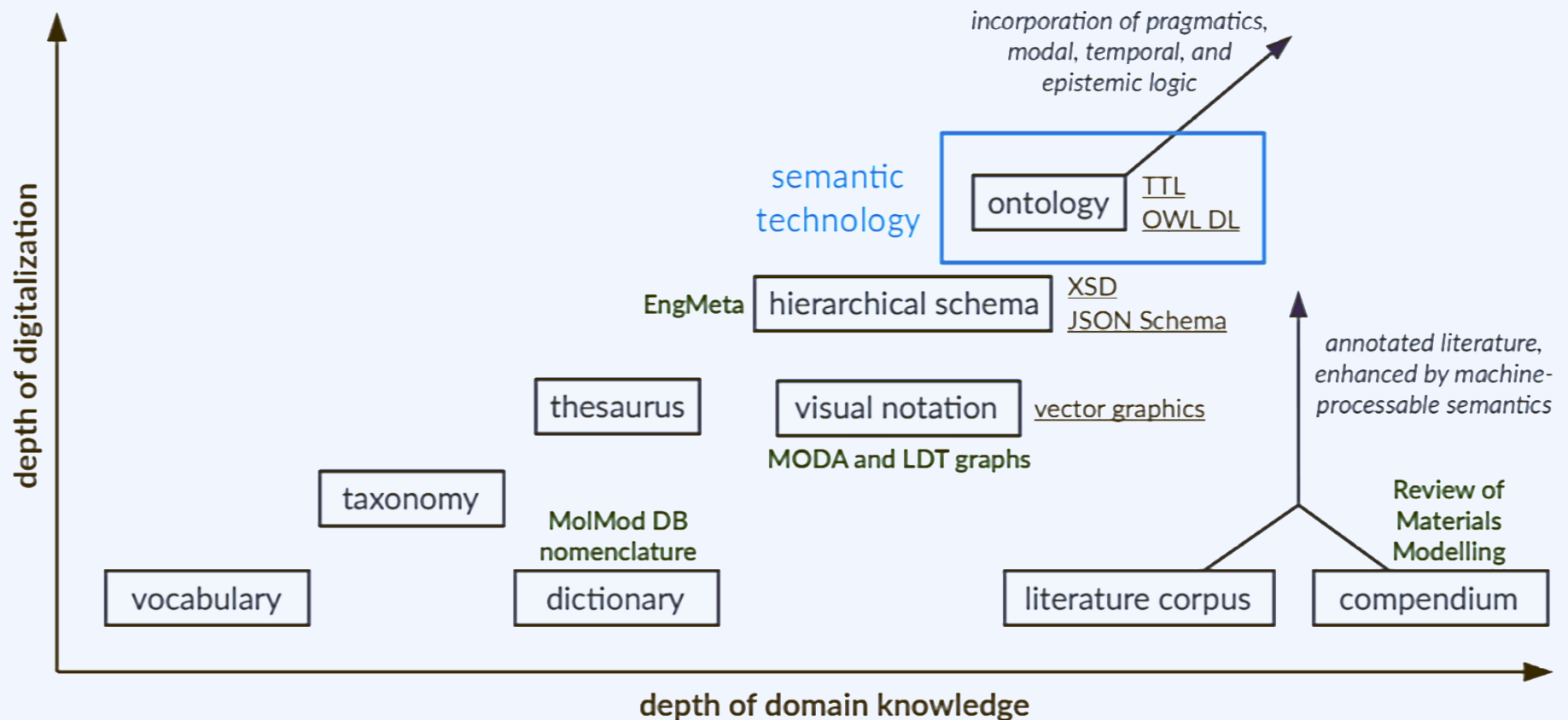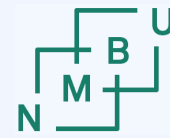
Persistent identifiers (PIDs)

The issuer of a PID must guarantee that it remains in place for the foreseeable future, and that metadata for the referenced object also remain in place.

Good PIDs are **GUPRIs**: Globally unique, persistent, resolvable identifiers.
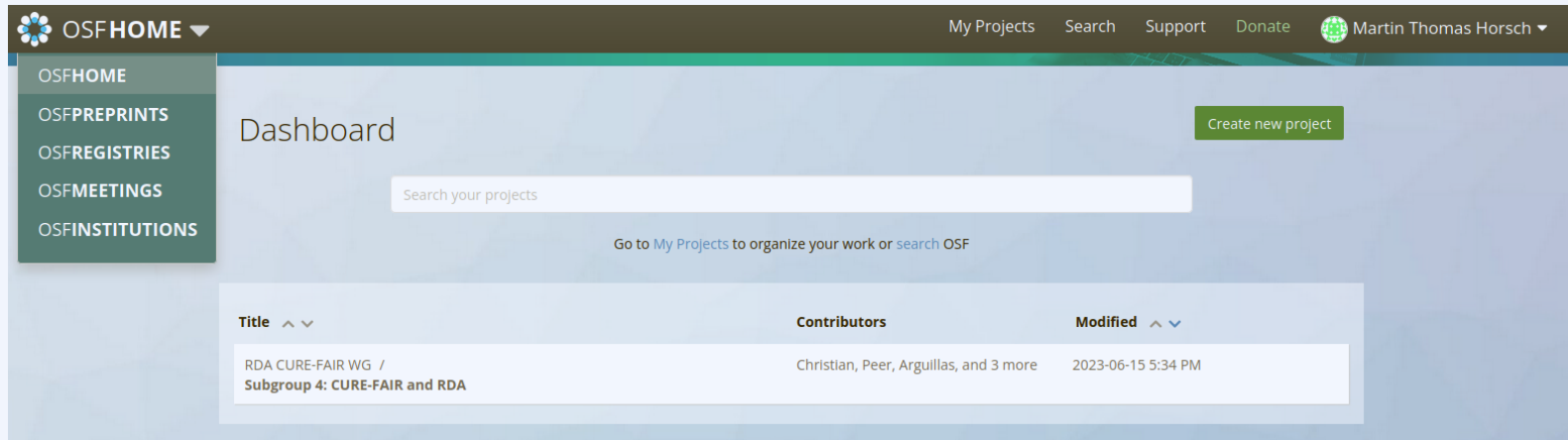
Examples:

- For researchers: **ORCID**.

- For books: **ISBN**. For journals: **ISSN**; however, these are not resolvable.

- For organizations: **ISNI** and **ROR**.

- Redirects to any digital artefacts, e.g., software development projects: Persistent URLs (**PURL**s); however, these are not globally unique.

- For documents, datasets, or any other digital objects: **DOI**.

Everybody engaged in research should have an ORCID.

Platforms for creating digital object identifiers
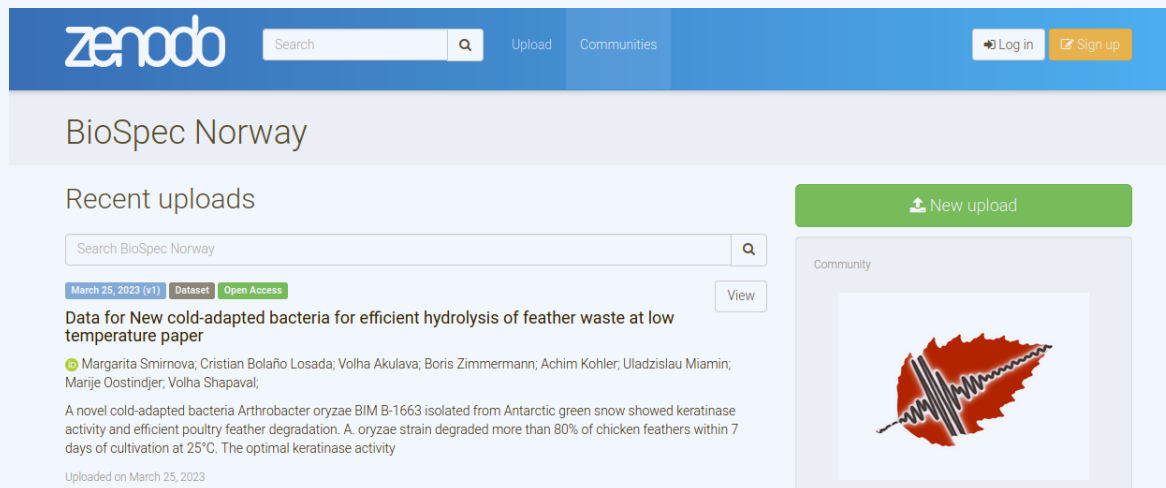
Open Science Framework (OSF) – see also the OSF hands-on guide

Zenodo

Check published data for FAIRness

Self-evaluation form issued by Direktoratet for e-helse[1]

F-UJI:[2] Automated FAIR data assessment tool

[1]FAIR evaluation form (Norwegian), Direktoratet for e-helse, **2020**.

[2]https://www.f-uji.net/

Good practices beyond FAIR

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

Reproducible Get the same answer asked of the same or different dataset

Interpretable Be clear, concise, accessible, and unambiguous

Open Open, inclusive, diverse, collective effort

Transparent Whenever possible, make public every part of research

https://riotscience.co.uk/
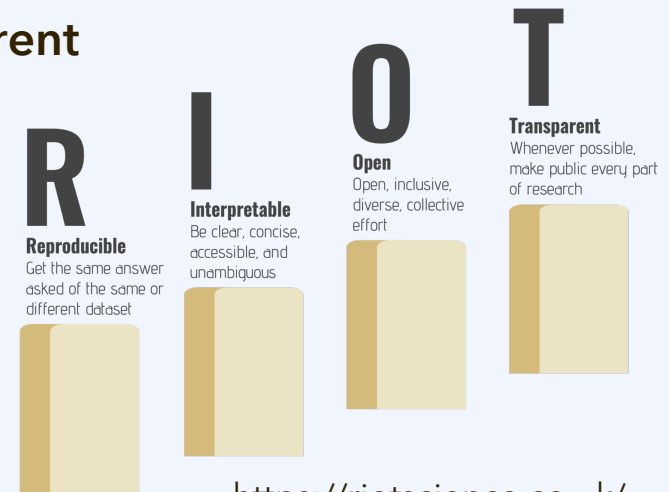
CARE:[2] Collective benefit, authority to control, responsibility, ethics
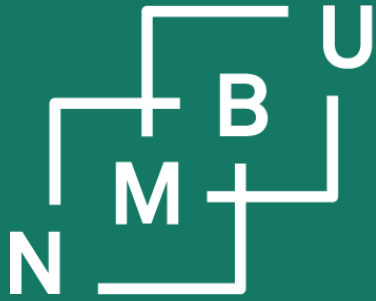
- Origin: Global Indigenous Data Alliance

- Uptake supported by the Research Data Alliance

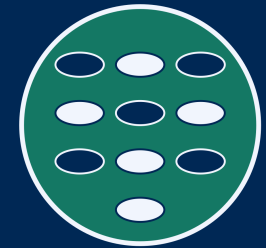- Orientation: Sovereignty and epistemic justice

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**.

DAT390

Data science seminar

3   Research methodology

3.1   Going beyond the state of the art

3.2   Research data management