TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Terminologies in Database Systems

Frühjahrstreffen 2024 der Fachgruppe Datenbanken

***Beyond silos**: Next steps in research data management*

Felix Engel

Jena, March 11.2024

Leibniz Association

Agenda

1. Metadata for Research Data Management
2. Terminology Service
3. Overcome Data Silos with Ontologies

nfdi Nationale Forschungsdaten Infrastruktur

NFDI4ing

NFDI4Ing Terminology Service

Metadata for Research Data Management

Intro

Research data is of inestimable value (*)

- *Empowers research*
- *Create innovation*

New research bases on existing research data (*)

- *Transparent*
- *Reproducible*

} *Reusable*

Current situation in Germany (*)

- *Data is stored decentraly*
- *Stored temporarily*
- ***Non-standardised metadata***
- *Varied quality*

***Nationale Forschungsdateninfrastruktur* (NFDI)** establishes an infrastructure for high qualitative RDM to foster research reusability

nfdi Nationale Forschungsdaten Infrastruktur

NFDi4ing

(*) https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

Metadata for Research Data Management

Current situation

- Commonalities between **metadata** in **databases** and in the **library context, like**
  - **Structural:** Structure, organizational aspects (e.g. relation: scientific publication and supplementary data)
  - **Descriptive:** Creator, and keywords, …
  - **Contextual Information:** Rights management, language, descriptions, …

- **Used** in **databases** to**:** organize, integrate, govern, analyse, …

- What about reproducibility? Contextual-**metadata**, can carry **reproducibility** relevant information

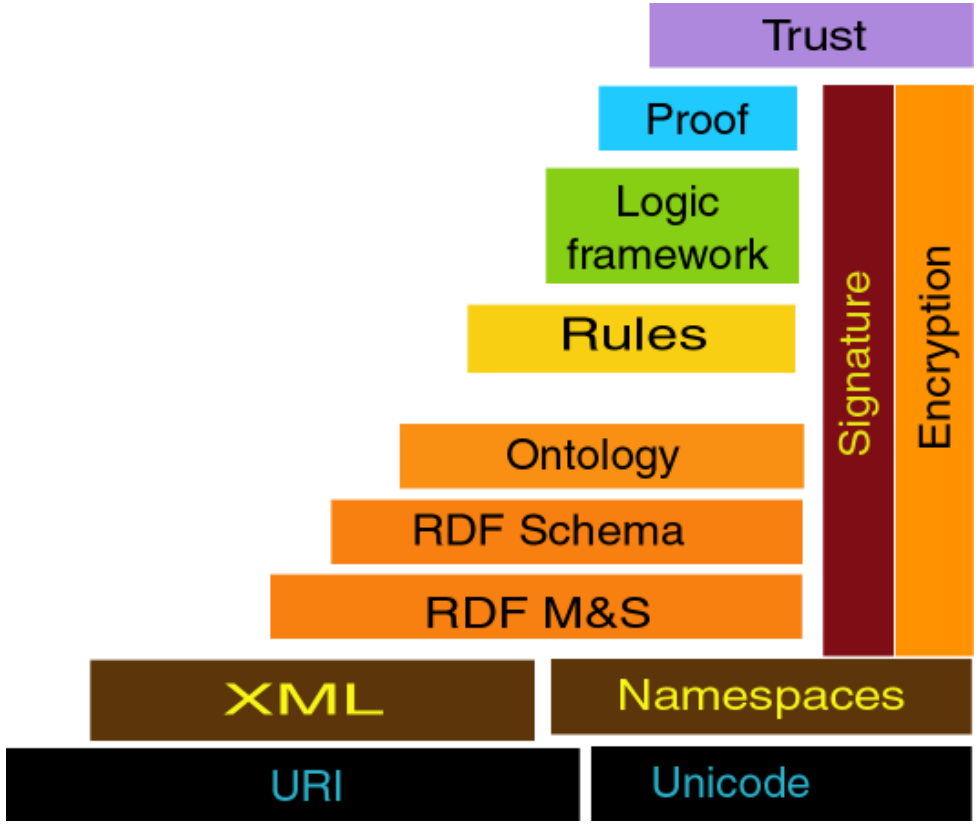
- … requires unambiguous, **community specific** terminologies and expressive formalisms
  - **Content indexing**. E.g. micro- and macronutrients [*]
    - Soil scientists: *nitrogen, phosphorus, potassium, calcium, magnesium and sulphur*
    - Nutritionists: *carbohydrates, protein and fat*
  - **Workflow descriptions**: wet lab experiment, data-driven, input, output, configuration ...

[*] Jordan, I., Heil, E., & Keding, G. (2021). Coming to terms with terminology in agriculture-nutrition research projects: an interactive glossary. *Ernährungsumschau*, *68*(10), 198-203.

Metadata for Research Data Management

Current situation

- More **expressive** options
  - **Thesaurus:** semantically grouped terms
    - Writing and editing: get synonyms, antonyms, …
    - Semantic text analysis: word definitions, …
    - Knowledge Organisation and classification: classification of documents, …
  - **Ontology:** formal representation of knowledge
    - Formalisation of knowledge: description of processes, …
    - Semantic Interoperability: shared and machine processable information, …
    - Decision support: inferencing, …



Semantic Web Stack.
Taken from: https://www.w3.org/2004/Talks/0611-sb-wsswintro/slide18-0.html

Metadata for Research Data Management

Challenges with Ontologies

- **Domain specific** (Engineering, Culture, Chemistry, …) and community specific
- **Evolving** continuously and dynamically over time
- Must be **accepted, developed and maintained** by a **designated community** (avoid isolated solution!). Includes i.e.
  - promotion (make community aware of its existence)
  - aligned with further metadata initiatives (moving away from silos)
  - applicable in RDM (in RDM practice)
- NFDI4Ing supports terminology development and use through introduction of engineering specific **Terminology Services** for Ontologies

NFDI4Ing Terminology Service

NFDI4ing

Terminology Service

- In general: Our **Terminology Service** is a **web based platform** that support take-up and standardisation of Ontologies

- **Utilised as a**
  - **Developer (e.g. Back-End Services):** Data- and Knowledge Management tasks. I.e.
    - Content indexation with controlled vocabularies
    - Search: Query reformulation, term suggestion, …
  - **Knowledge Engineers.** I.e.
    - Bundles ontologies of a domain
    - Provides meta data and statistics
    - Search for and within ontologies
    - Makes alignments visible
  - **Authors as dissemination point,** fosters awareness and alignment

NFDI4Ing Terminology Service

NFDI4ing

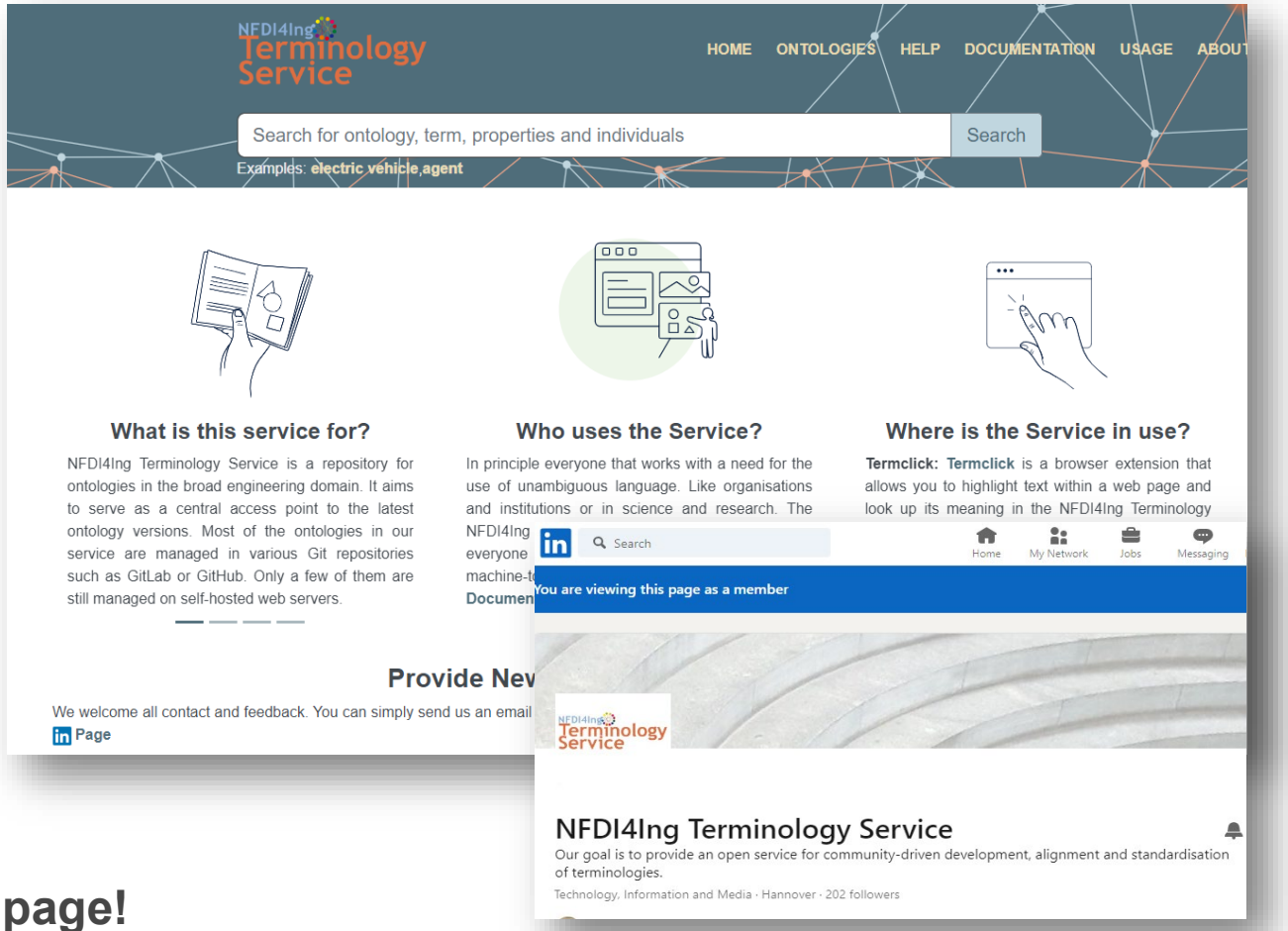
Terminology Service

NFDI4Ing Terminology Service

- **NFDI4Ing TS**: https://terminology.nfdi4ing.de/ts/

- **Some statistics**
  - 93 ontologies
  - ~160.000 Terms
  - ~12.000 properties

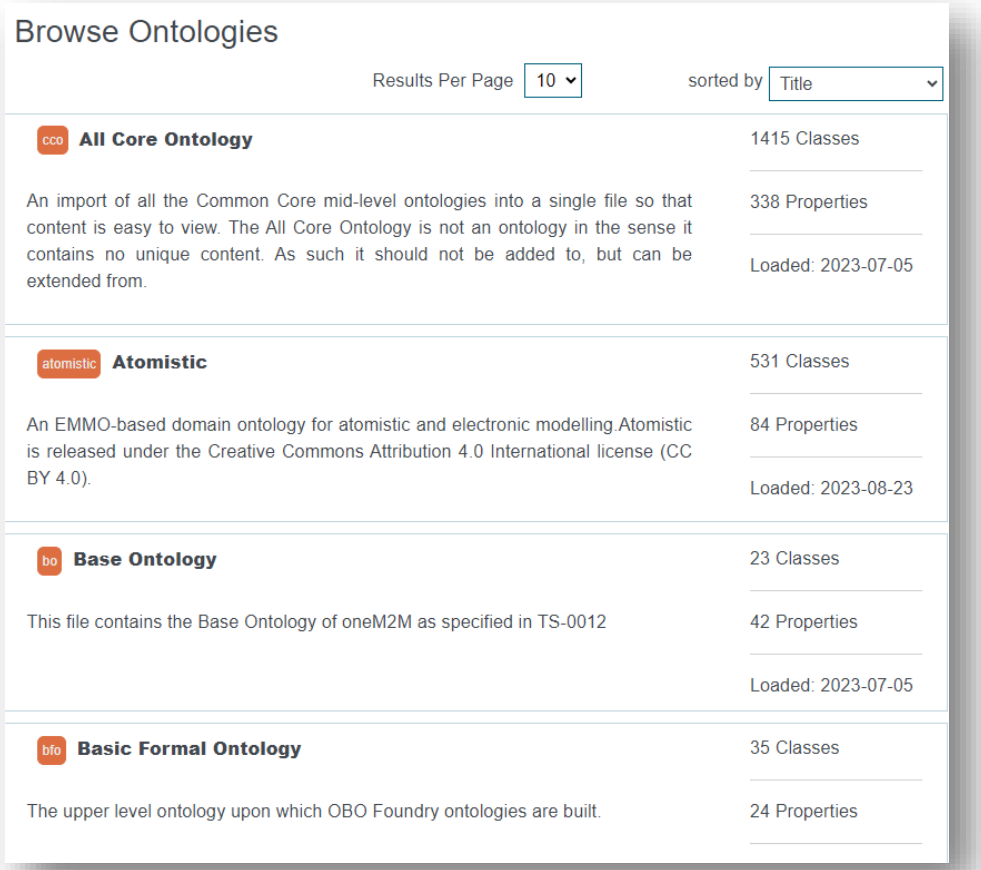
- **Functional service offer**
  - Free text search searching
    (for- and within ontologies)
  - Browsing and Filtering (by various metadata)
  - Visualisation
  - Issue tracker
  - Machine to machine communication
    (REST interfaces)

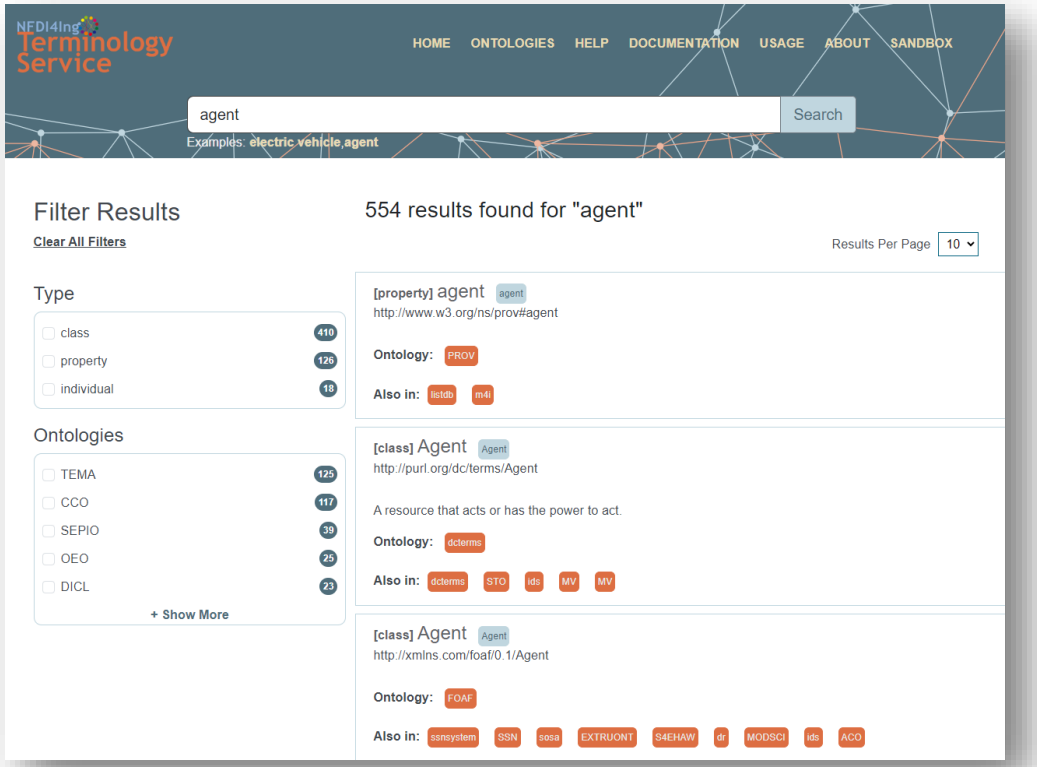


Stay up to date with our LinkedIn page!

Terminology Service

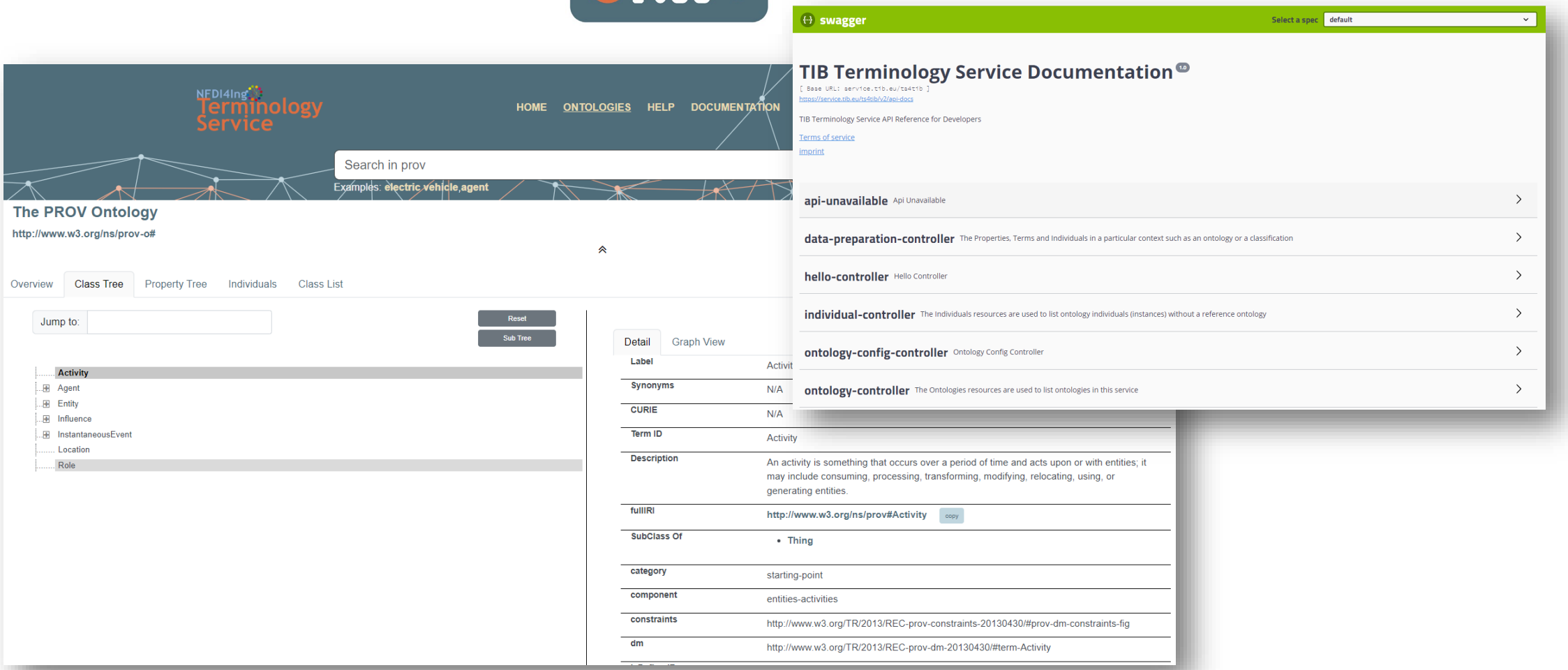
NFDI4Ing Terminology Service





Terminology Service

NFDI4Ing Terminology Service



Overcome Data Silos with Ontologies

DM applications based on de facto standard ontologies

Simple Experiment: *Semantification of Space Data – A feasibility Study*

- **Wikidata ontology:** https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology
- **SpaCy** pre-trained NER and Linking model

- **Data:** Planetary Data System 4 is a long term **archive** for **data products** from NASA's planetary missions

- **Approach:** Create **Knowledge Graph** and apply **federated SPARQL query**
  - Apply existing **SpaCy** model: entity extraction and linking (**Wikidata**)
  - **Federation:** PDS4 Knowledge Graph - Wikidata

- **Manually evaluated**

Query term	Record count	Concepts
Planet	20	Mars (14), Planet (5), Jupiter (1), Pluto (1), Neptune (1)
Solar System	20	same as 'Planet'
Superior Planet	20	same as 'Planet'
Vehicle	15	Spacecraft (15)
Vehicle*	20	Spacecraft (15), Lander (5)
Planetary Probe	5	Lander (5)

2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)

Overcome Data Silos with Ontologies

DM applications based on de facto standard ontologies

astrophysics data system

- Follow-up experiment**:** Train supervised **ML model to label text** with Ontology terms

- De facto standard: **Unified Astronomy Thesaurus** (UAT)
  - > 2000 concepts
  - Polyhierarchy
  - 11 UAT Top concepts as Entity Types

- **Text categorisation**
  - **Named Entity Recognition** of **UAT** top concepts (11)
  - Train empty **SpaCy NER** model
  - **First results**
    - Standard configuration
    - Precision: 80.86, Recall: 85.06 , F1: 82.65
    - Good, but needs improvement (distortion through “Others”)

- Kindly received **training data** from NASA ***astrophysics data system*** (ads)

spaCy

WIKIDATA

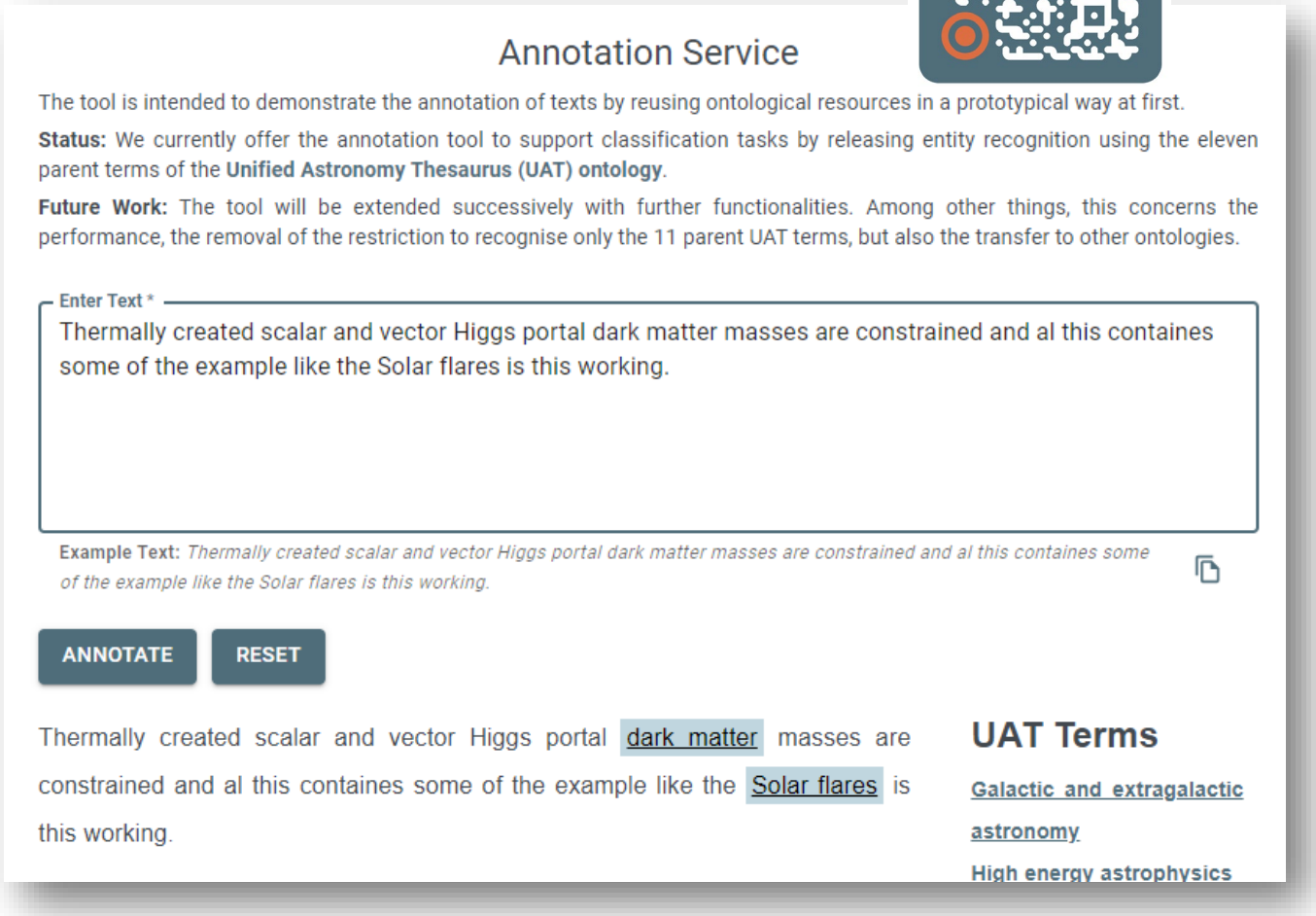
Astrophysical processes	66.08	75.35	70.41
Cosmology	67.91	62.24	65.06
Exoplanet astronomy	51.10	80.78	62.60
Galactic and extragalactic astronomy	55.00	64.65	59.44
High energy astrophysics	65.45	57.86	61.42
Interdisciplinary astronomy	71.97	71.92	71.95
Interstellar medium	62.98	83.84	71.92
Observational astronomy	69.83	76.06	72.81
Solar physics	57.87	72.34	64.30
Solar system astronomy	69.78	72.89	71.30
Stellar astronomy	62.21	61.53	61.87
Other	98.18	99.23	98.66
Total	**80.86**	**85.05**	**82.65**

Overcome Data Silos with Ontologies

DM applications based on de facto standard ontologies

- **Positive example:**
  *Thermally created scalar and vector Higgs portal **dark matter** masses are constrained and al this containes some of the example like the **Solar flares** is this working.*

- **Negative** examples:
  *she recalled the thrill of exploring the abandoned **Mars** factory with her friends*



Overcome Data Silos with Ontologies

DM applications based on de facto standard ontologies

NFDI4Ing Terminology Service

- **Challenge:** find words on the tip of the tongue
  - *Fear of spiders* → Arachnophobia
  - *Integrated Circuits* → Wafer or Chipset

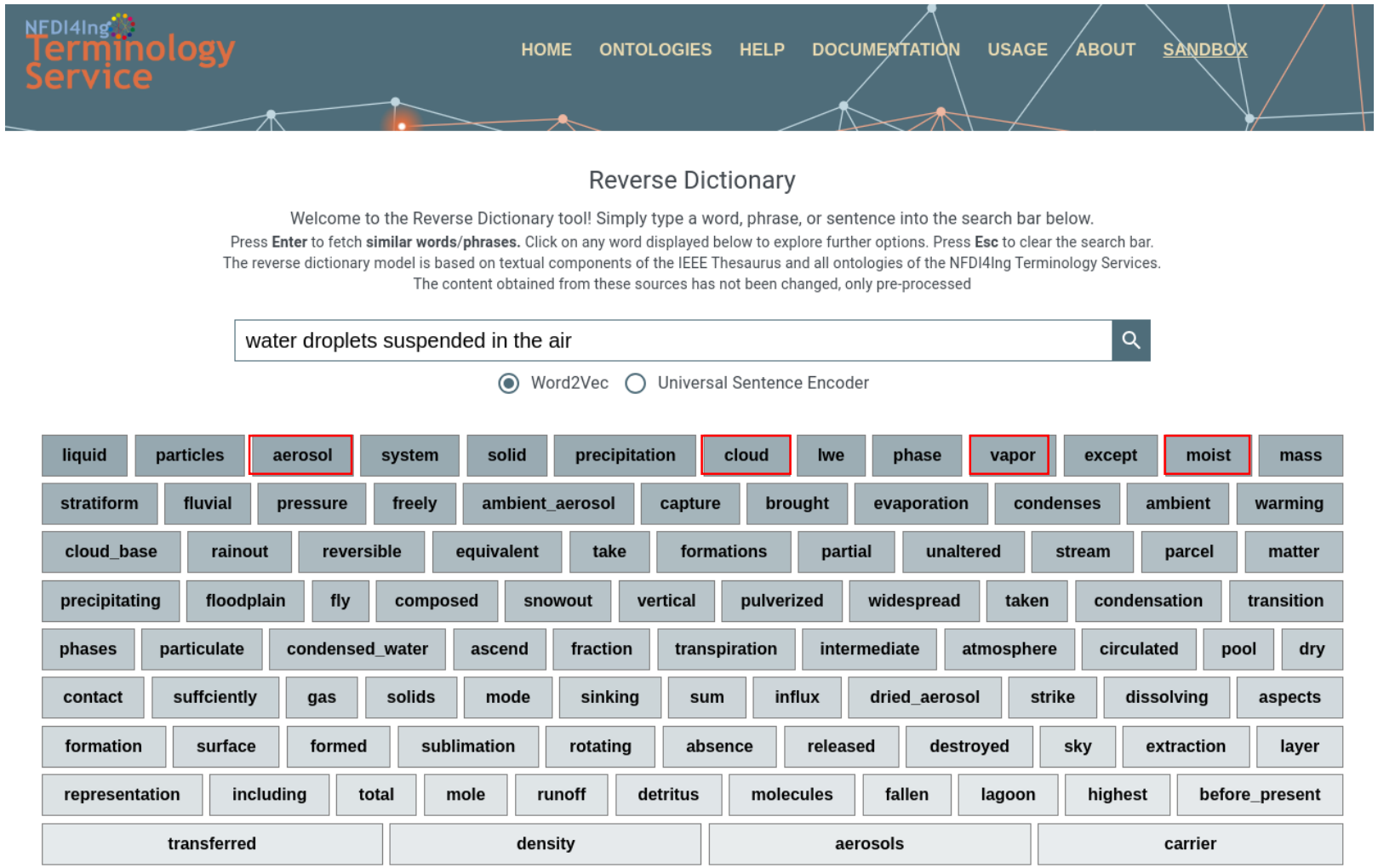
- **Hypothesis:** Domain ontologies encode commonly accepted terminology for unambiguous information exchange, useful to train a Reverse Dictionary

Dictionary vs. **Reverse Dictionary**

Look up words alphabetically to find their definitions

Start with a concept or an idea and try to find the word that best represents it

Overcome Data Silos with Ontologies



NFDI4Ing Terminology Service

Earth System Sciences

Query: *water droplets suspended in the air*

Overcome Data Silos with Ontologies

Pros and Cons

- **Pros**
  - Means to standardize: **overcome language barriers**
  - Model very **complex interdependencies** (reproducibility)
  - Very **active field** research area

- **Cons**
  - **Availability** is depending on domain
  - **Hard to understand** as domain expert
  - Still **not enough tooling** available

THANKS