#EOSCsymposium23

Semantic interoperability for data and metadata

21 September | 10.40 - 11.40

eosc

EOSC Future

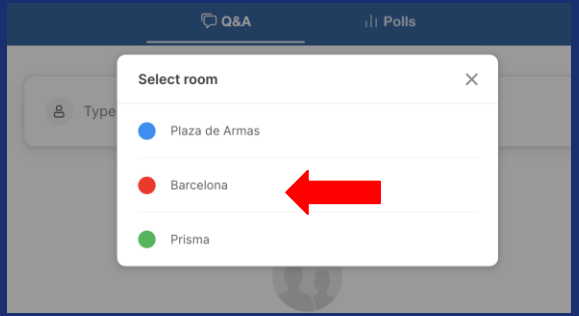
EU23 SPANISH PRESIDENCY COUNCIL OF THE EUROPEAN UNION

Funded by the European Union

Join the discussion on Slido.com

Use the code: EOSCsymposium23 or the QR code

Select the room: Barcelona



eosc

EOSC Future

EU23 SPANISH PRESIDENCY COUNCIL OF THE EUROPEAN UNION

Funded by the European Union

Slido.com -> Enter code: EOSCsymposium23
Select room: Barcelona

Opening and introduction to the session

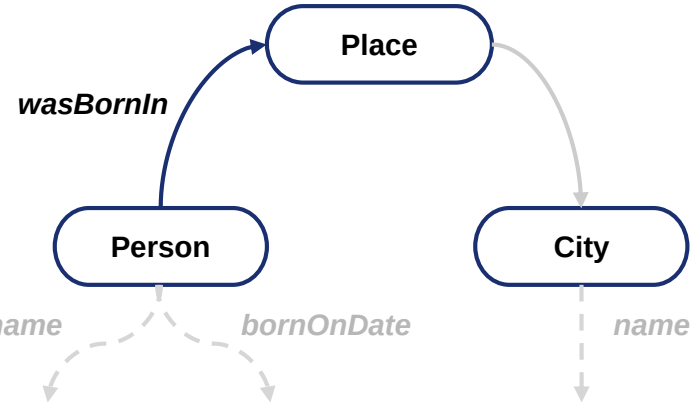
Presented by Wolmar Nyberg Åkerström

Room:
Barcelona

A web of FAIR data and services for research

“What is sent is what is understood”

- Semantic models connect data with relevant research concepts
- Used to translate and exchange information to support a variety of research related use cases
- Interoperability across tools, workflows and infrastructures



people	birth dates	birth place
Bob Dylan	1941-05-24	Duluth
Elvis Presley	1935-01-08	Tupelo
Little Richard	1932-12-05	Macon

“We are EOSC”: A call to action for all of us

“From Gutenberg to Berners–Lee”

- Roadmap & priorities in the Strategic Research and Innovation Agenda (SRIA) for EOSC document
- Practices & skills
- Standards, tools & services
- Federated infrastructure

“[...] **invest in the creation, adoption and governance** of community-based metadata and data standards [...]”

“[...] semantic artefact catalogues in **national infrastructures and guidelines**”

“[...] **support for publishing** semantic artefacts through institutional or vocabulary specific thematic repositories”

Making sense of your group's data

" data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings"

How about another group's data?

Making sense of another group's data

“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

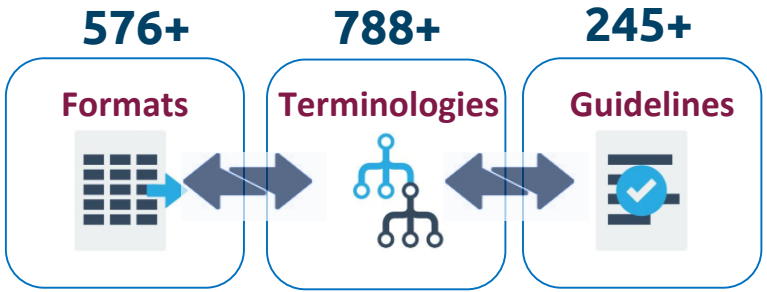
“ data should be assessable so that judgments can be made about their reliability and the competence of those who created them”.

How about Europe's research data?

Making sense of Europe's research data

" data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings"

" data should be assessable so that judgments can be made about their reliability and the competence of those who created them".



Data storage & exchange

Requirements & checklists

Conceptual models & controlled vocabularies

* Numbers from the FAIRsharing Registry

Connecting the web of FAIR data and services

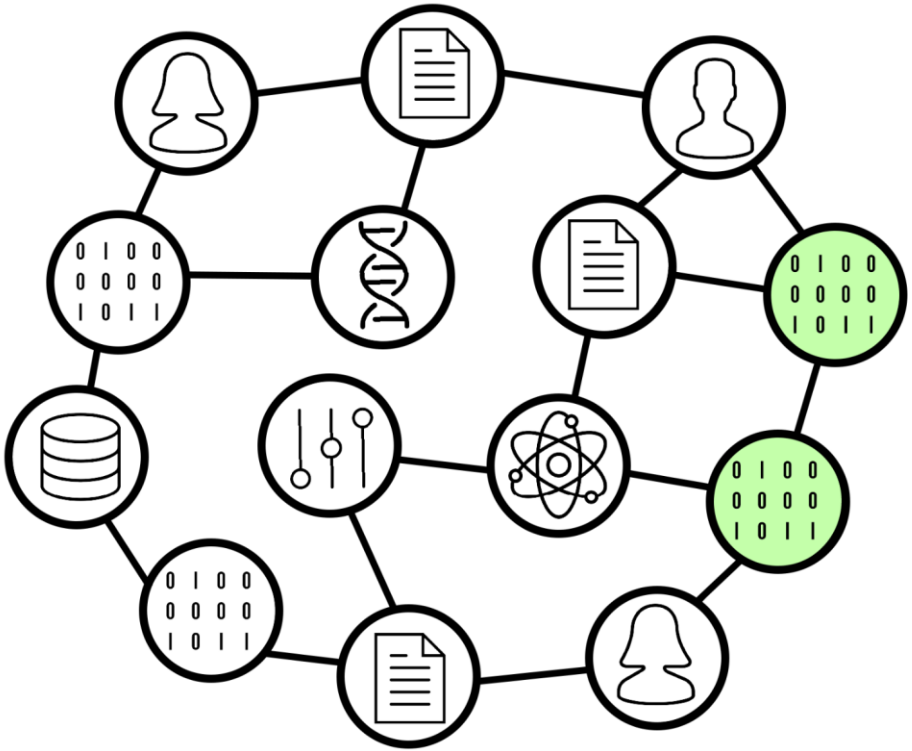
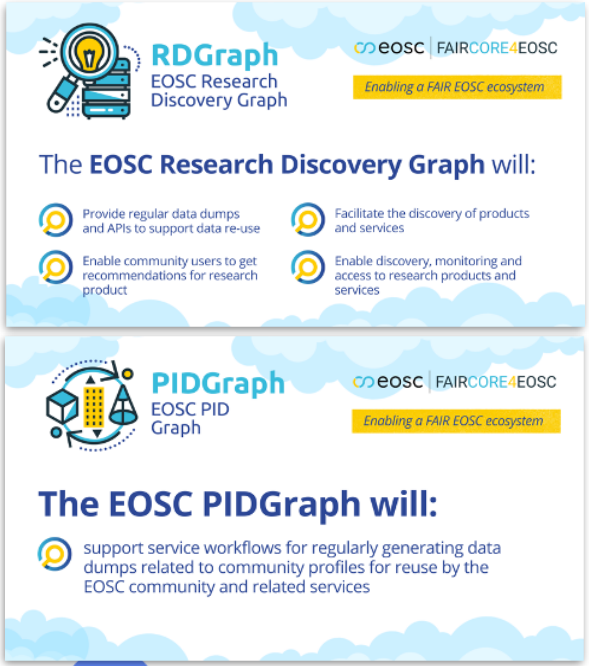
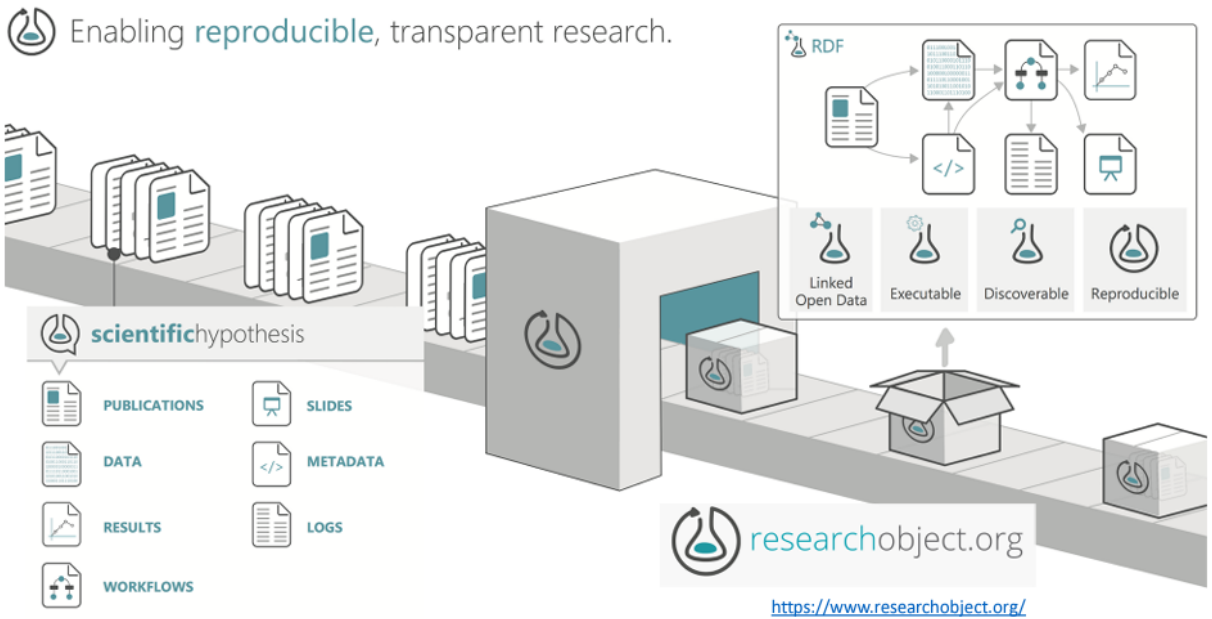


Illustration: “PID Graph” from the FREYA project

Highlight: FAIRCORE4EOSC



Leveraging communities and consensus



https://www.researchobject.org/

Highlight: EOSC Association

Semantic interoperability for data and metadata

Thursday 21 September

10:30-11:30 CEST

Room: Barcelona

8 min **Opening and introduction to the session**
Wolmar Nyberg Åkerström

8 min **Semantic artefacts and their representations**
Yann Le Franc

8 min **Catalogues of semantic artefacts and their governance**
Susanna-Assunta Sansone

8 min **Mappings, crosswalks and alignment**
Daan Broeder

8 min **Implementation examples**
Alexandra Kokkinaki

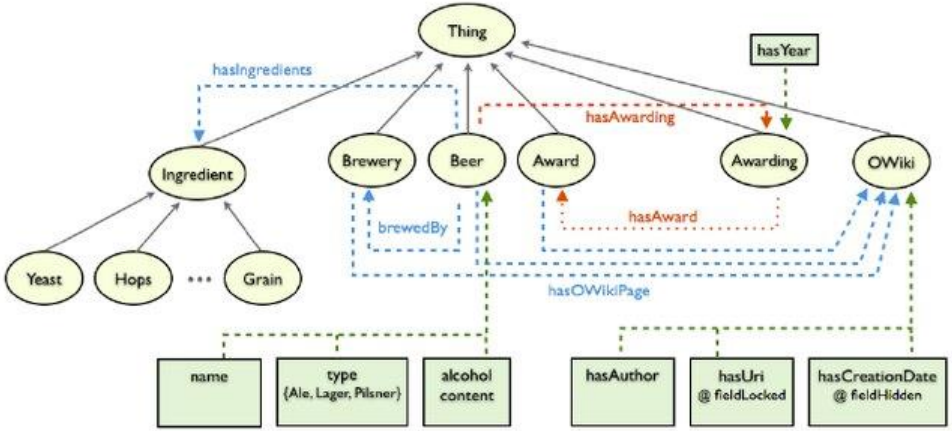
20 min **Q&A and Panel discussion: Opportunities to promote and converge on best practices**

Semantic artefacts and their representations

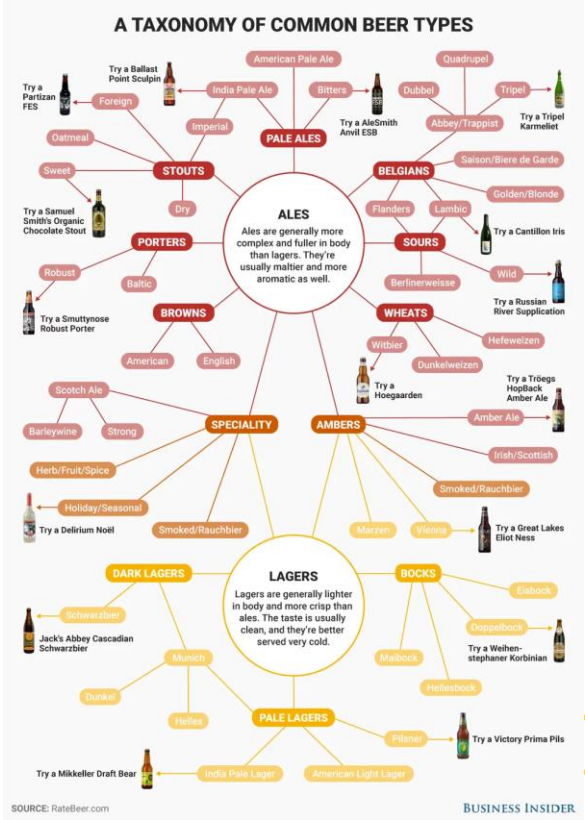
Presented by Yann Le Franc

What do semantic artefacts represent?

Conceptual models of the world which describes the concepts and their relations necessary to describe data and its context.



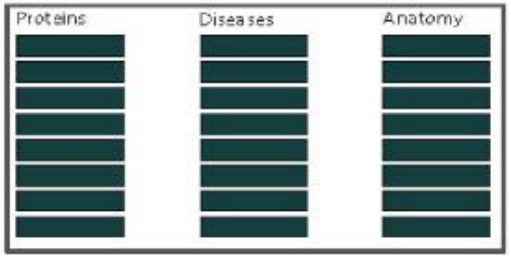
From Di Iorio, Angelo & Peroni, Silvio & Vitali, Fabio. (2012). OWiki: Enabling an Ontology-Led Creation of Semantic Data. 10.1007/978-3-642-23172-8_24.



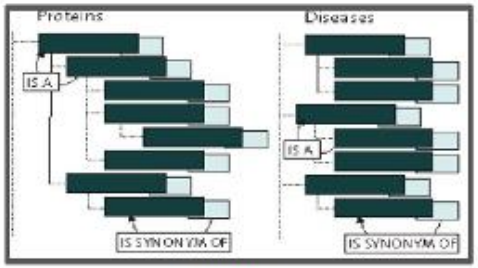
from https://www.businessinsider.com/different-types-of-beer-2016-13?r=US&IR=T

The diversity of Semantic artefacts

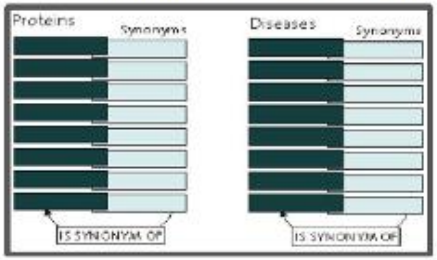
Controlled Vocabulary (CV):
An authoritative list of terms



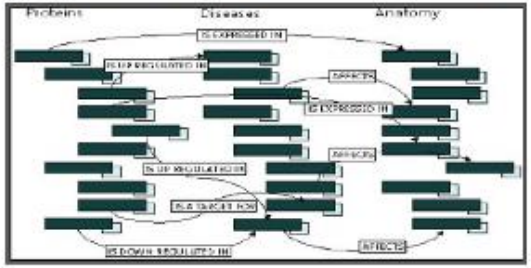
Taxonomy:
A CV with a tree-hierarchical (parent/child term) structure



Thesaurus:
A kind of taxonomy with structure and specific types of relationships between terms



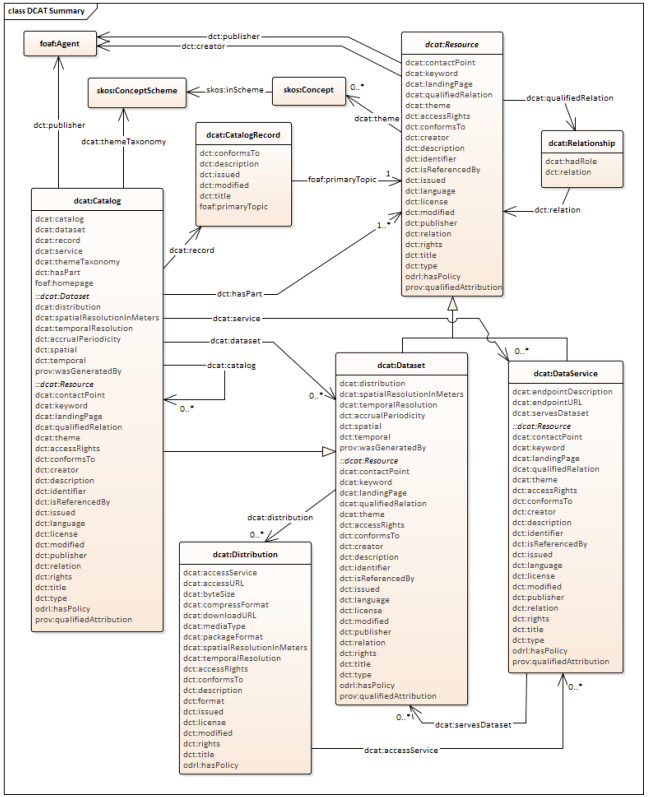
Ontology:
is a kind of taxonomy, but the types of relationships are greater in number and more specific in their function



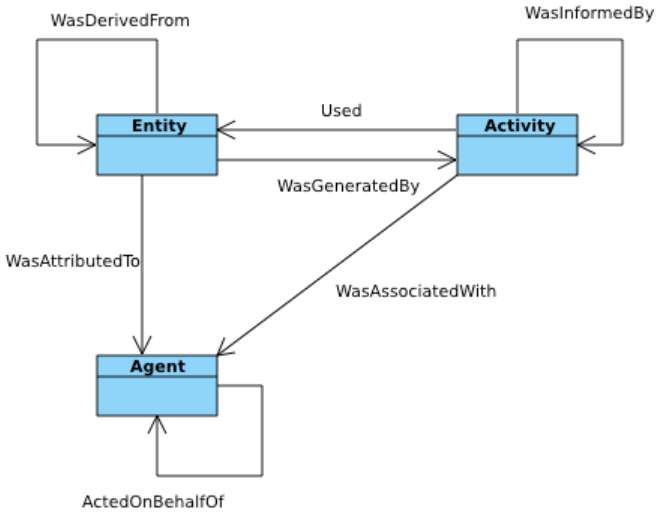
From: "Terminology for FAIR data competency – Considerations: before we get our hands dirty!" (2019), S.A. Sansone and P. McQuilton – Terminology Workshop – May 20-21 2019 - Paris

Examples of generic conceptual models

DCAT

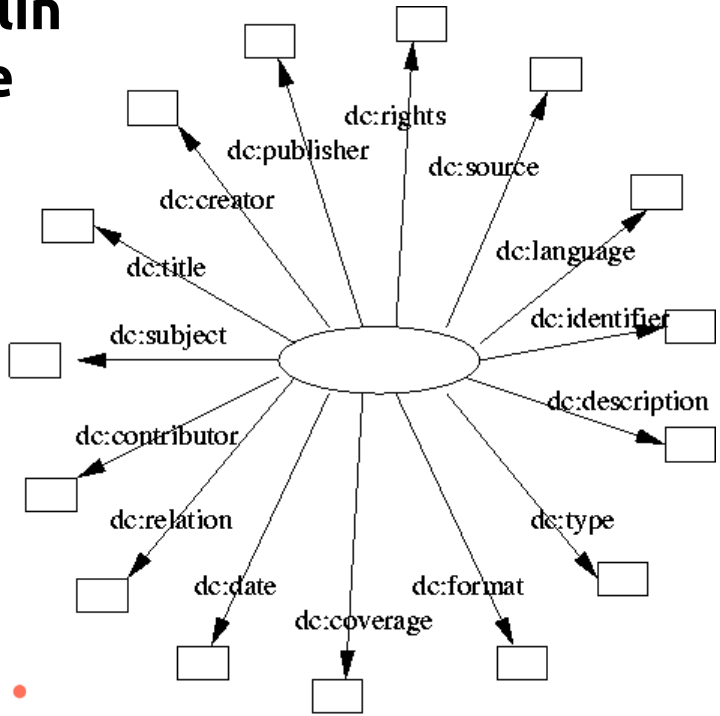


PROV



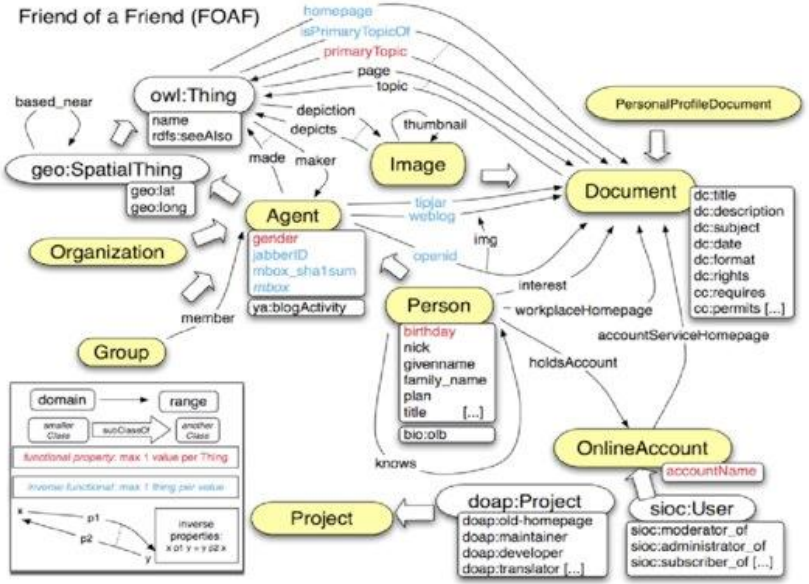
Examples of generic conceptual models

Dublin Core



https://www.dublincore.org/specifications/dublin-core/dcq-rdf-xml/2001-08-29/

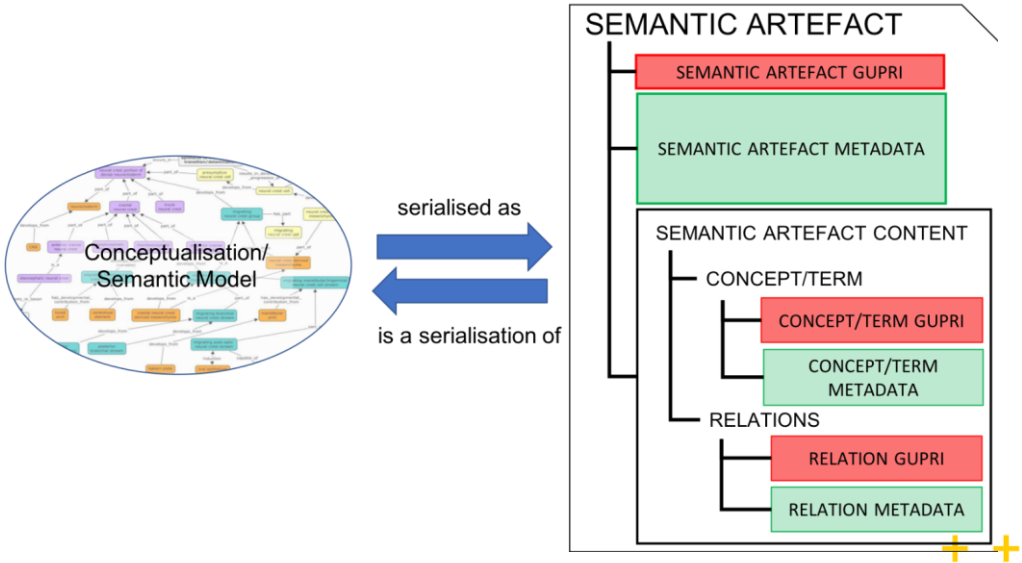
FOAF



Challenger, Moharram. (2012). The Ontology and Architecture for an Academic Social Network. International Journal of Computer Science Issues. 9.

What are semantic artefacts?

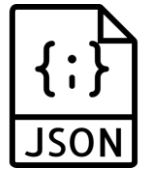
"A ***semantic artefact*** is defined within our work as a machine-actionable and -readable formalisation of a conceptualisation, enabling sharing and reuse by humans and machines. These artefacts may have a broad range of formalisation, from loose sets of terms, taxonomies, thesauri to higher-order logics. Moreover, semantic artefacts are serialised using a variety of digital representation formats …"- Yann Le Franc, Luiz Bonino, Hanna Koivula, Jessica Parland-von Essen, & Robert Pergl. (2022). D2.8 FAIR Semantics Recommendations Third Iteration (V1.0). Zenodo. https://doi.org/10.5281/zenodo.6675295



FAIRSFAIR
Fostering Fair Data Practices In Europe

What are the common representations?

A wide diversity of digital representation of the conceptual models.



Examples of practices in various communities

Biodiversity and environmental science

GBIF Repository of Schemas

GBIF Resources

Name	Last modified	Size
Parent Directory		-
distribution_status/	2022-02-22 11:35	-
establishment_means/	2022-02-22 11:35	-
occurrence_status/	2022-02-22 11:35	-
taxonomicStatus/	2022-02-22 11:35	-
agent_role.xml	2020-05-12 17:22	11K
dataset_subtype.xml	2020-05-12 17:22	6.6K
dataset_type.xml	2020-05-12 17:22	3.2K
dataset_type_2015-07-10.xml	2020-05-12 17:22	3.8K
description_type.xml	2020-05-12 17:22	15K
direction.xml	2020-05-12 17:22	6.2K
distribution_status_2020-07-15.xml	2020-07-15 10:32	4.5K
establishment_means.xml	2020-05-12 17:22	3.3K
life_form.xml	2020-05-12 17:22	4.0K
life_stage.xml	2020-05-12 17:22	12K
nomenclatural_code.xml	2020-05-12 17:22	3.8K
nomenclatural_status.xml	2020-05-12 17:22	18K
nomenclatural_status_2019-02-08.xml	2020-05-12 17:22	18K
occurrence_status.xml	2020-05-12 17:22	4.5K
occurrence_status_2020-07-15.xml	2020-07-15 10:32	1.3K
preservation_method.xml	2020-05-12 17:22	6.7K
quantity_type_2015-07-10.xml	2020-05-12 17:22	7.0K
quantity_type_2023-02-28.xml	2023-02-28 13:36	7.5K
rank.xml	2020-05-12 17:22	20K
rank_2015-04-24.xml	2020-05-12 17:22	25K
reference_type.xml	2020-05-12 17:22	3.8K
resource_type.xml	2020-05-12 17:22	4.6K

<xml />

Examples of practices in various communities

Biodiversity and environmental science

GBIF Repository of Schemas

Name	Last modified	Size
Parent Directory		-
distribution_status/	2022-02-22 11:35	-
establishment_means/	2022-02-22 11:35	-
occurrence_status/	2022-02-22 11:35	-
taxonomicStatus/	2022-02-22 11:35	-
agent_role.xml	2020-05-12 17:22	11K
dataset_subtype.xml	2020-05-12 17:22	6.6K
dataset_type.xml	2020-05-12 17:22	3.2K
dataset_type_2015-07-10.xml	2020-05-12 17:22	3.8K
description_type.xml	2020-05-12 17:22	15K
direction.xml	2020-05-12 17:22	6.2K
distribution_status_2020-07-15.xml	2020-07-15 10:32	4.5K
establishment_means.xml	2020-05-12 17:22	3.3K
life_form.xml	2020-05-12 17:22	4.0K
life_stage.xml	2020-05-12 17:22	12K
nomenclatural_code.xml	2020-05-12 17:22	3.8K
nomenclatural_status.xml	2020-05-12 17:22	18K
nomenclatural_status_2019-02-08.xml	2020-05-12 17:22	18K
occurrence_status.xml	2020-05-12 17:22	4.5K
occurrence_status_2020-07-15.xml	2020-07-15 10:32	1.3K
preservation_method.xml	2020-05-12 17:22	6.7K
quantity_type_2015-07-10.xml	2020-05-12 17:22	7.0K
quantity_type_2023-02-28.xml	2023-02-28 13:36	7.5K
rank.xml	2020-05-12 17:22	20K
rank_2015-04-24.xml	2020-05-12 17:22	25K
reference_type.xml	2020-05-12 17:22	3.8K
resource_type.xml	2020-05-12 17:22	4.6K

eLTER Vocabularies

Vocabularies About Help | auf Deutsch

EnvThes - Thesaurus for long term ecological research, monitoring and experiments

Content language English Search

Alphabetical Hierarchy

A B C D E F G H I J K L M N O

P Q R S T U V W X Y Z M !* 0-9

A horizon
A layer → A horizon
abiotic environment
abiotic heterogeneity
ABL → atmospheric boundary layer
above ground
above ground biomass
above ground net primary production
above-ground net primary production
aboveground net primary production → above ground net primary production
aboveground production
absorbed
absorbed dose
absorbed dose rate
absorption
absorption coefficient
absorptive coefficient → absorption coefficient
absorptivity coefficient → absorption coefficient
abundance
abundance estimate
abundance of animal species → animal species abundance
abundance of bacillariophyceae → bacillariophyceae abundance
abundance of bacteria → bacteria abundance

Vocabulary information

TITLE	EnvThes - Thesaurus for long term ecological research, monitoring and experiments Environmental Thesaurus
DESCRIPTION	Thesaurus for long term ecological research, monitoring, experiments
CREATOR	https://orcid.org/0000-0003-2195-3997
RIGHTS	https://spdx.org/licenses/CC0-1.0
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
URI	http://vocabs.lter-europe.net/EnvThes/

Download this vocabulary: TURTLE

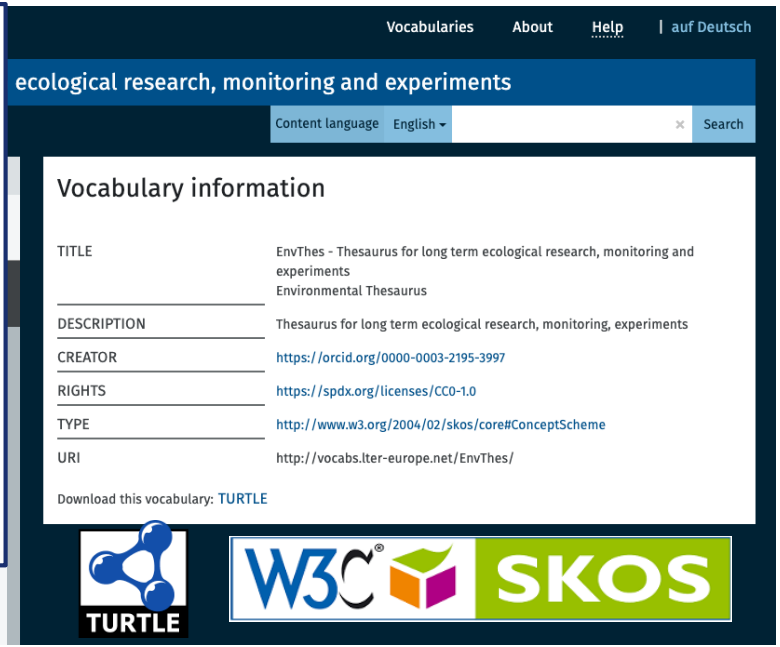
TURTLE

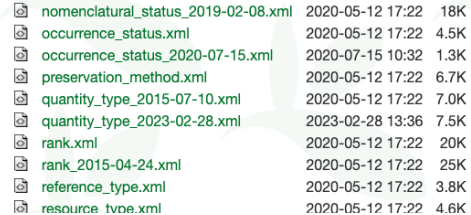
W3C SKOS

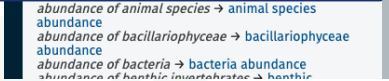
<xml />

Examples of practices in various communities

Biodiversity and environmental science



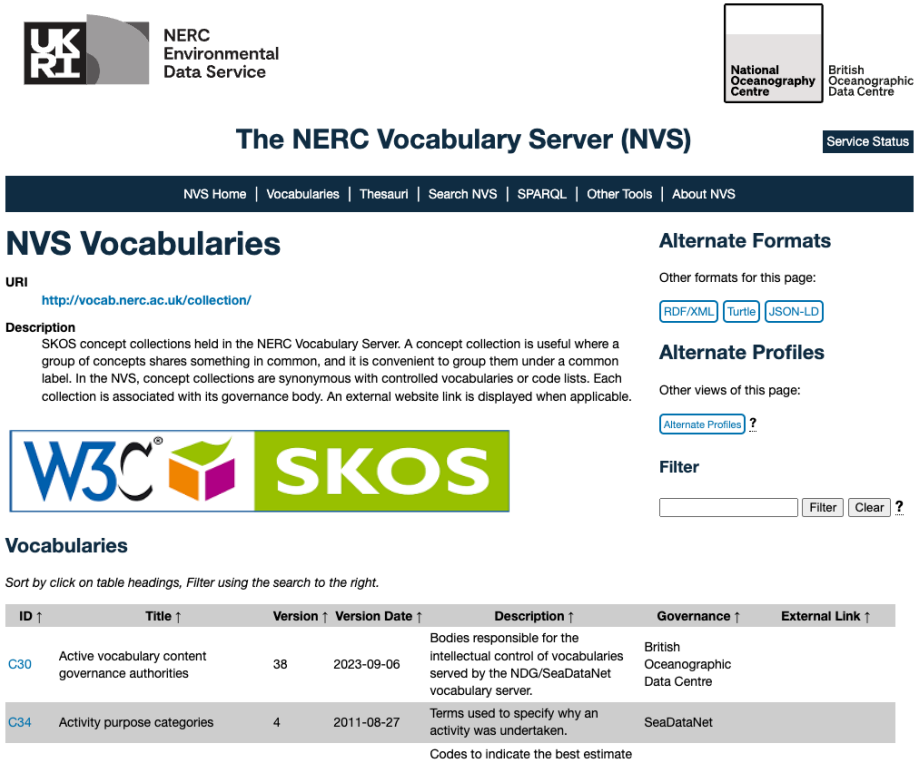




<xml />

Examples of practices in various communities

Marine, Agriculture, Food, Biomedical science



NERC Environmental Data Service

National Oceanography Centre — British Oceanographic Data Centre

The NERC Vocabulary Server (NVS)

Service Status

NVS Home | Vocabularies | Thesauri | Search NVS | SPARQL | Other Tools | About NVS

NVS Vocabularies

URI

http://vocab.nerc.ac.uk/collection/

Description

SKOS concept collections held in the NERC Vocabulary Server. A concept collection is useful where a group of concepts shares something in common, and it is convenient to group them under a common label. In the NVS, concept collections are synonymous with controlled vocabularies or code lists. Each collection is associated with its governance body. An external website link is displayed when applicable.

W3C SKOS

Alternate Formats

Other formats for this page:

RDF/XML | Turtle | JSON-LD

Alternate Profiles

Other views of this page:

Alternate Profiles ?

Filter

Filter | Clear ?

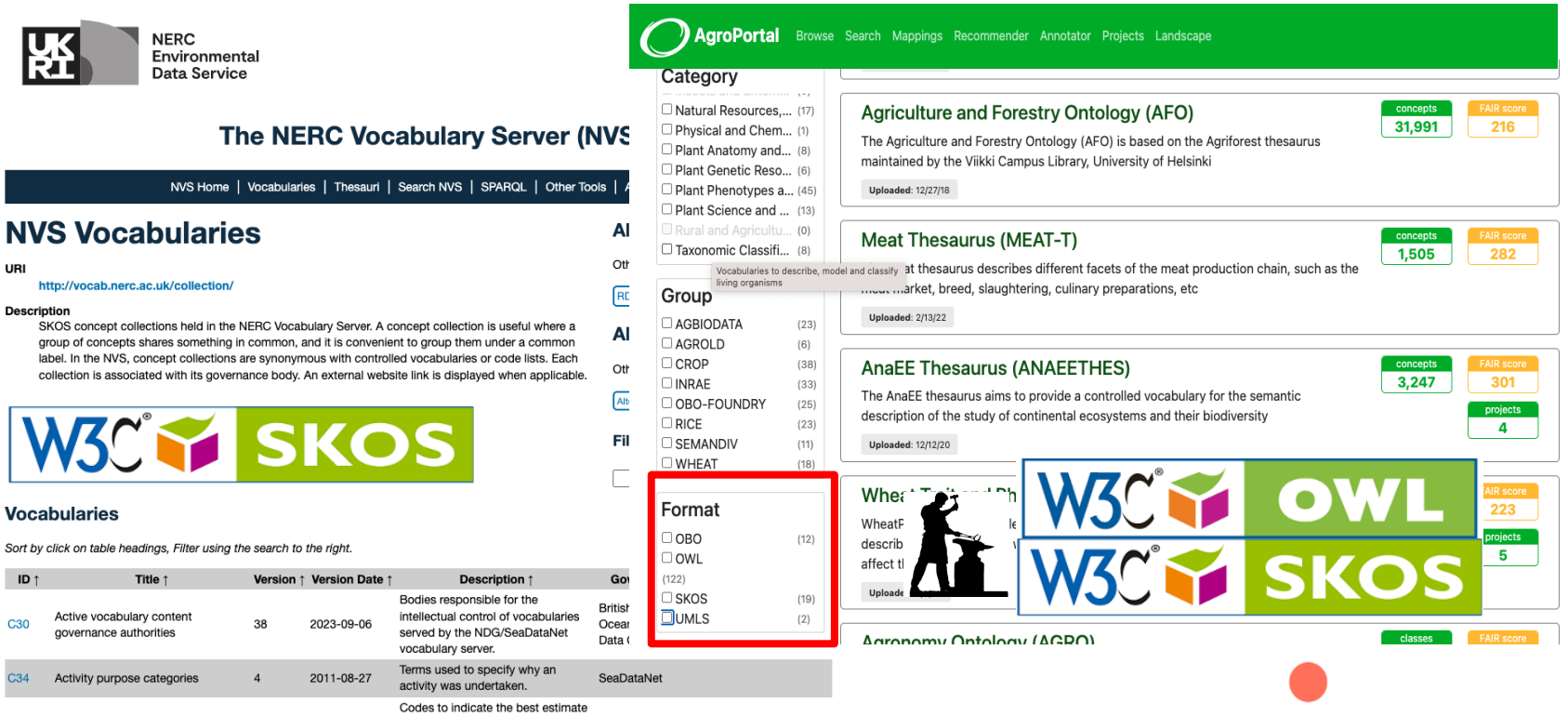
Vocabularies

Sort by click on table headings, Filter using the search to the right.

ID ↑	Title ↑	Version ↑	Version Date ↑	Description ↑	Governance ↑	External Link ↑
C30	Active vocabulary content governance authorities	38	2023-09-06	Bodies responsible for the intellectual control of vocabularies served by the NDG/SeaDataNet vocabulary server.	British Oceanographic Data Centre	
C34	Activity purpose categories	4	2011-08-27	Terms used to specify why an activity was undertaken.	SeaDataNet	
				Codes to indicate the best estimate		

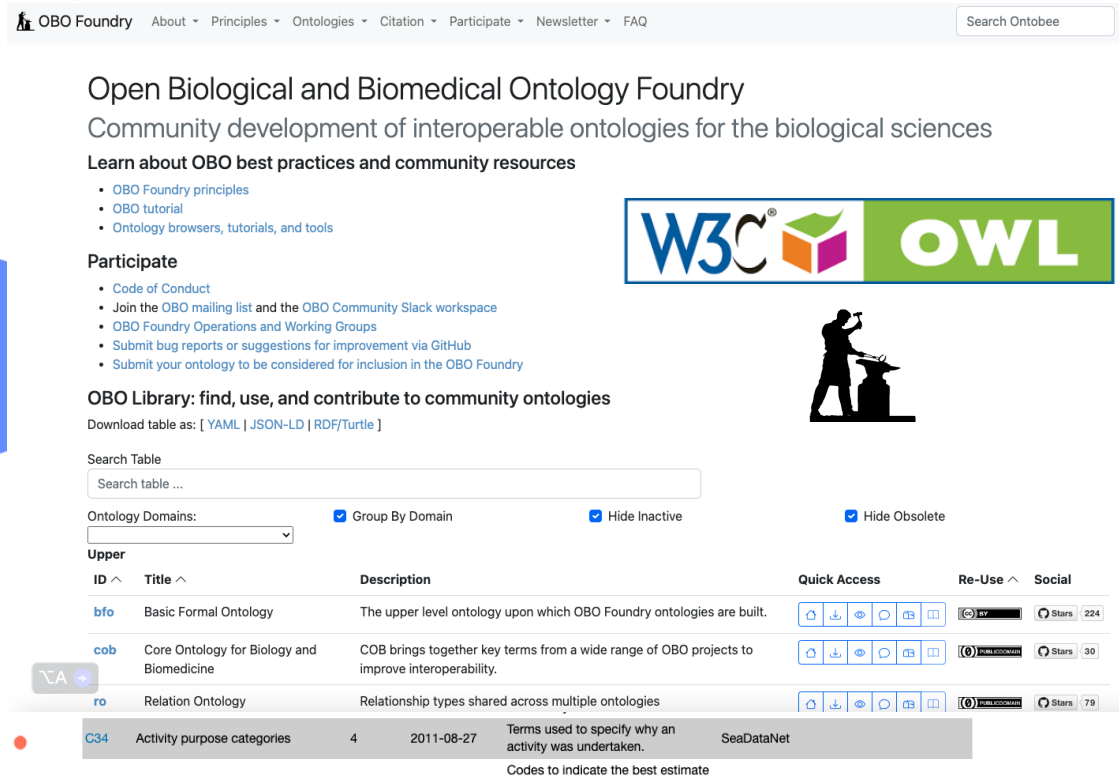
Examples of practices in various communities

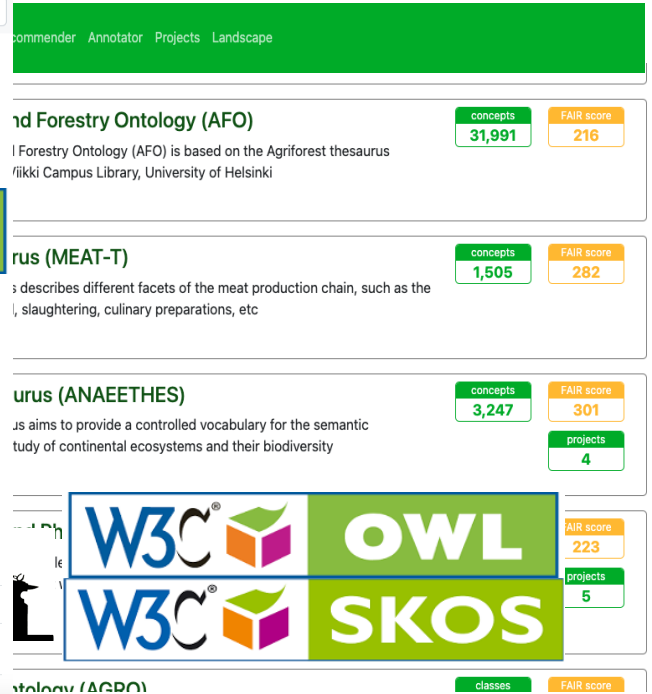
Marine, Agriculture, Food, Biomedical science



Examples of practices in various communities

Marine, Agriculture, Food, Biomedical science





Examples of practices in various communities

Marine, Agriculture, Food, Biomedical science

OBO Foundry About Principles Ontologies Citation Participate Newsletter FAQ

Open Biological and Biomedical Ontology Foundry

Community development of interoperable ontologies for the

Learn about OBO best practices and community resources

- OBO Foundry principles
- OBO tutorial
- Ontology browsers, tutorials, and tools

Participate

- Code of Conduct
- Join the OBO mailing list and the OBO Community Slack workspace
- OBO Foundry Operations and Working Groups
- Submit bug reports or suggestions for improvement via GitHub
- Submit your ontology to be considered for inclusion in the OBO Foundry

OBO Library: find, use, and contribute to community ontologies

Download table as: [YAML | JSON-LD | RDF/Turtle]

Search Table

Search table ...

Ontology Domains: Group By Domain Hide Inactive

Upper

ID	Title	Description
bfo	Basic Formal Ontology	The upper level ontology upon which OBO Foundry ontologies are buil
cob	Core Ontology for Biology and Biomedicine	COB brings together key terms from a wide range of OBO projects to improve interoperability.
ro	Relation Ontology	Relationship types shared across multiple ontologies

C34	Activity purpose categories	4	2011-08-27	Terms used to specify why an activity was undertaken.	SeaDat

Codes to indicate the best estimate

Category

- All Organisms (44)
- Anatomy (78)
- Animal Development (13)
- Animal Gross Anatomy ...
- Arabidopsis (2)
- Biological Process (58)

Group

- BIBLIO (10)
- BIS (3)
- CGIAR (1)
- CTSA (6)
- OBO_Foundry (11)
- PSI (4)

Format

- OBO (103)
- OWL (786)
- SKOS (103)
- UMLS (31)

BioPortal Ontologies Search Annotator Recommender Mappings

RxNORM (RXNORM) — projects 7, classes 107,690

RxNorm Vocabulary

Uploaded: 8/7/23

National Drug Data File (NDDF) — projects 1, classes 31,518

National Drug Data File Plus Source Vocabulary

Uploaded: 8/7/23

Read Codes, Clinical Terms Version 3 (CTV3) (RCD) — projects 2, classes 140,065

Clinical Terms Version 3 (CTV3) (Read Codes) (Q199): National Health Service National Coding and Classification Centre

Uploaded: 8/7/23

Foundational Model of Anatomy (FMA) — instances 2, classes 104,721, projects 17

FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy

Uploaded: 5/13/19

Medica — classes 352,426, notes 2

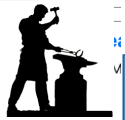
Medical S

Uploaded:

Logical Observation Identifier Names and Codes (LOINC) — projects 8, classes 285,529

Logical Observation Identifier Names and Codes (LOINC)

Uploaded: 8/7/23



Examples of practices in various communities

Linguistics, Humanities

Examples of practices in various communities

Linguistics, Humanities

CLAVAS: vocabular

CLARIN
CLAVAS

introduction
vocabularies
usage
API
FAQ
publications
contact
acknowledgements
change log

Introduction

CMDI 1.2 adds the ability to link extern

Vocabularies

Currently the following vocabularies are

ISO 639-3 (hdl:11459/CLAVAS_810f8d
Codes for the representation of n
of languages as possible, includir
with the CMDI Component iso-63

And the following vocabularies are beir

licenses
organizations
media types (also known as MIME type

Consortium of European Social Science Data Archives

Account

cessda
VS Vocabulary Service

Find Controlled Vocabulary

English (en)

Home | About | User Guide

33 Vocabularies

Results per page 30

Sort by CV name (A-Z)

Agency

Search agencies

gesis

Central concepts of R|EX research *(rexconcepts)*

This CV, developed within the framework of the DP-R|EX joint project, maps the central concepts and theoretical approaches in research on racism and right-wing extremism. The compilation is based on a systematic evaluation of the relevant national and international empirical research literature. The CV equally takes into account the different thematic (racism, right-wing extremism, discrimination) as well as methodological (qualitative research, standardised surveys, data from social media and messaging services) research strands.

Version: 1.2.0 Download

DE EN

DDI

Aggregation Method *(AggregationMethod)*

Identifies the type of aggregation used to combine r … information at a broader level than the level at which detailed observations are tak…

Version: 1.1.2 Download

EN DA NO

W3C SKOS

DDI

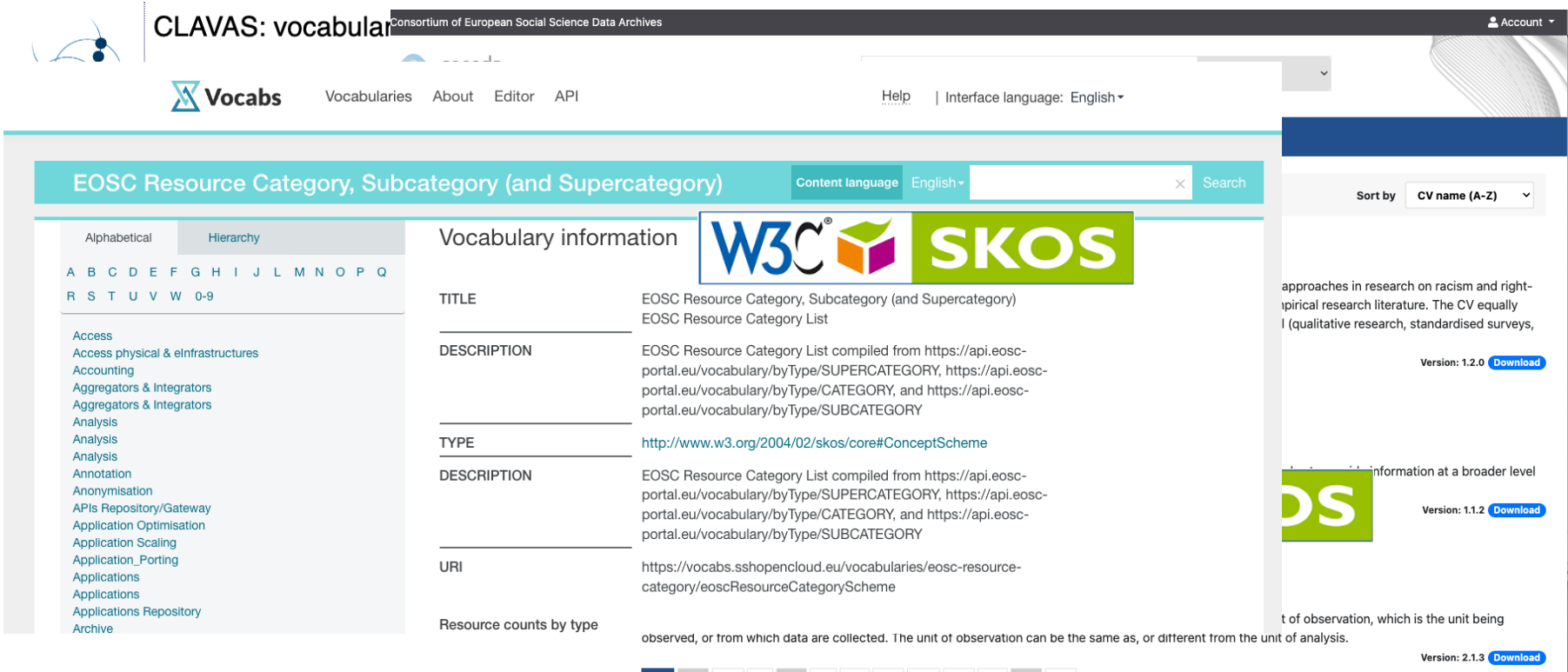
Analysis Unit *(AnalysisUnit)*

Describes the entity being analyzed in the study or variable. This vocabulary can also be used to describe the unit of observation, which is the unit being observed, or from which data are collected. The unit of observation can be the same as, or different from the unit of analysis.

Version: 2.1.3 Download

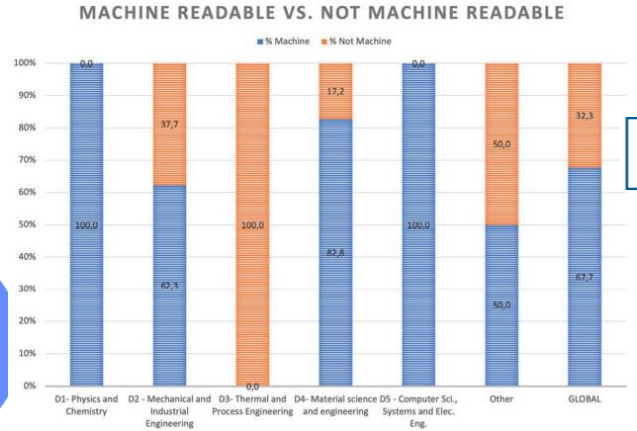
Examples of practices in various communities

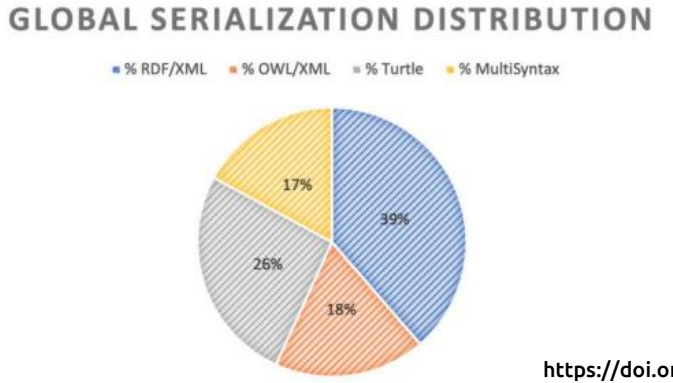
Linguistics, Humanities

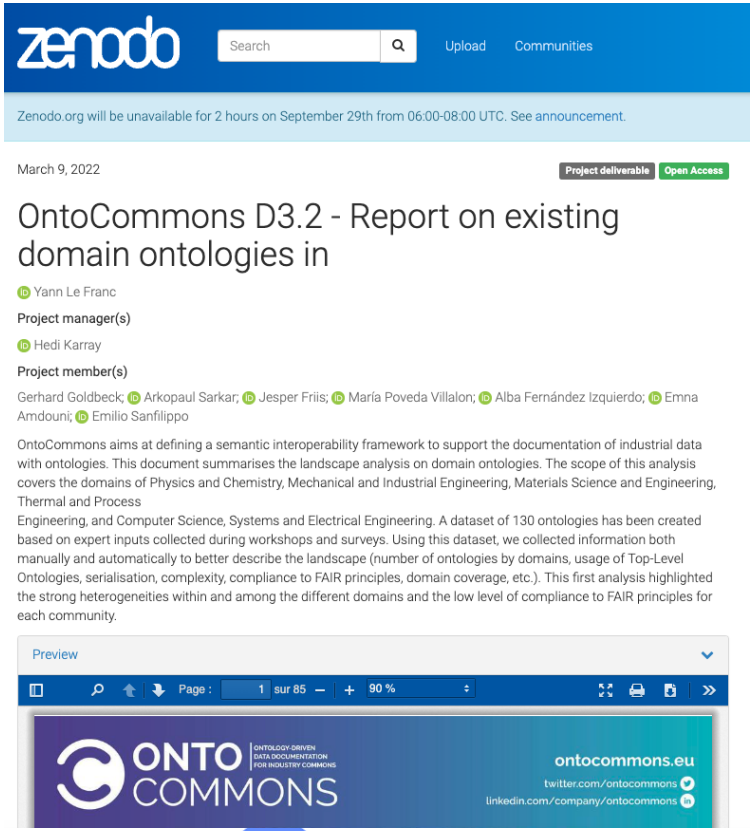


Examples of practices in various communities

Material Science and Manufacturing

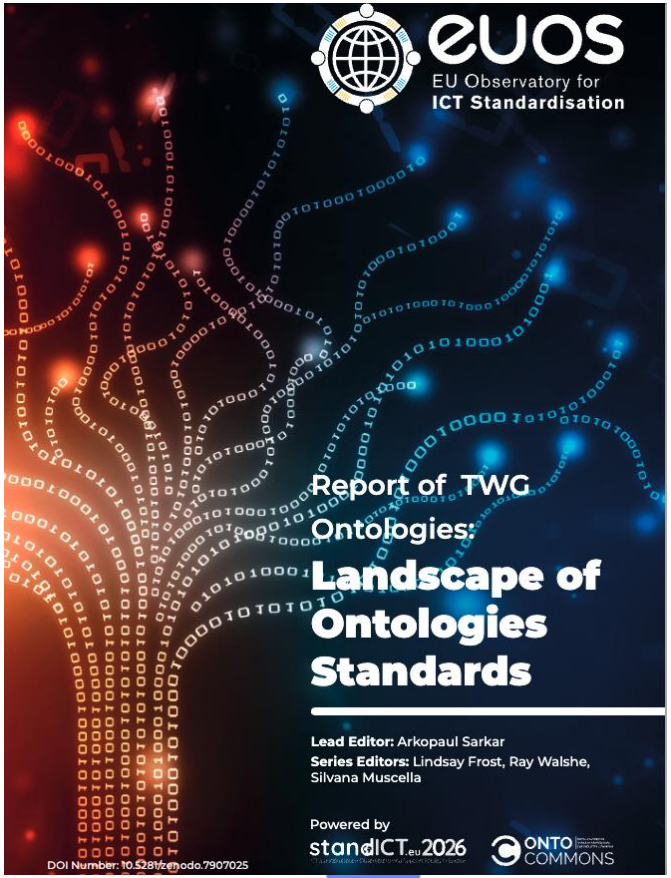






https://doi.org/10.5281/zenodo.6504553

Semantic artefacts as part of standardisation landscape

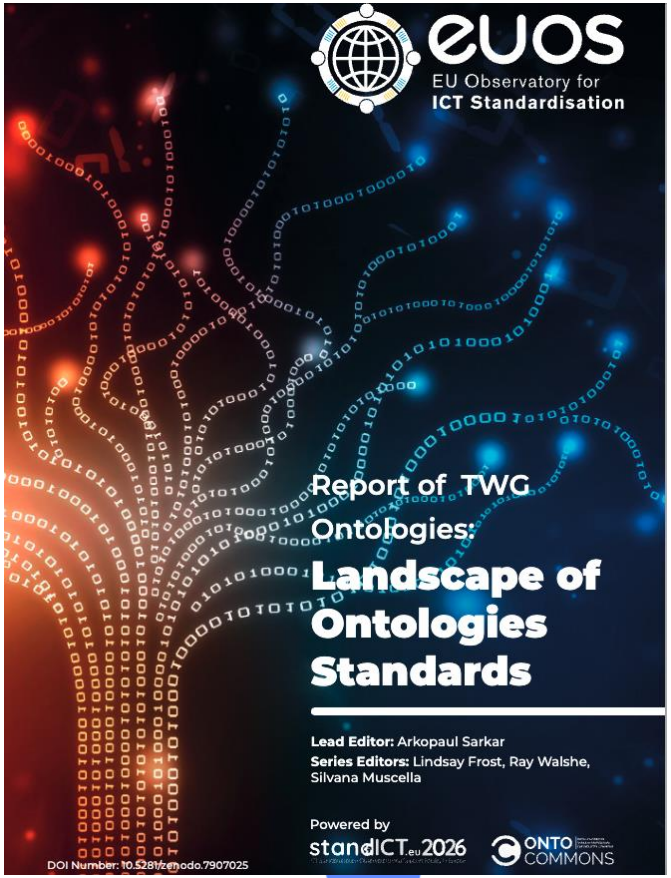


StandICT.eu, Sarkar, Arkopaul, Frost, Lindsay, Walshe, Ray, & Muscella, Silvana. (2023). Report of TWG Ontologies: Landscape of Ontologies Standards. Zenodo. https://doi.org/10.5281/zenodo.7907025

Semantic artefacts as part of standardisation landscape

Large diversity of Semantic Artefact representations, coupled with difficulty to find, interoperate and reuse which impair semantic interoperability

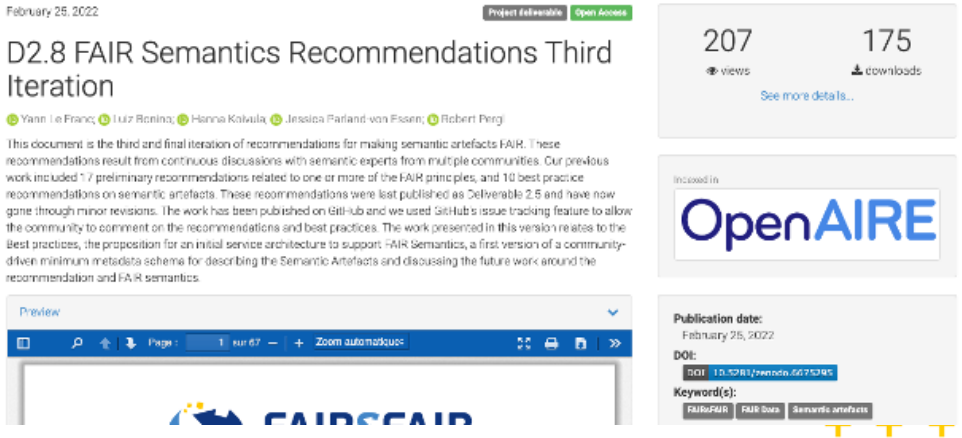
How can we resolve this situation?



StandICT.eu, Sarkar, Arkopaul, Frost, Lindsay, Walshe, Ray, & Muscella, Silvana. (2023). Report of TWG Ontologies: Landscape of Ontologies Standards. Zenodo. https://doi.org/10.5281/zenodo.7907025

Applying the FAIR principles to semantic artefacts

- 17 generic recommendations and 12 Best Practices
- Recommendation aligned with RFC 2119 (MUST, SHOULD, MAY)
  - 9 MUST
  - 7 SHOULD
  - 1 MAY
  - 1 Undetermined
- Minimum metadata profile for FAIR Semantic Artefacts



Clement Jonquet, Biswanath Dutta, Luiz O. Bonino da Silva Santos, Robert Pergl, Yann Le Franc. Common Minimum Metadata for FAIR Semantic Artefacts. 2nd Workshop on Ontologies for FAIR and FAIR Ontologies (Onto4FAIR), Cassia Trojahn; Luiz Olavo Bonino da Silva Santos; Giancarlo Guizzardi; Clement Jonquet, Jul 2023, Sherbrooke, Canada. ⟨hal-04106533v2⟩

https://doi.org/10.5281/zenodo.6675295

Summary of the recommendations

- Usage of globally unique persistent and resolvable identifiers for semantic artefacts, their content (i.e. concept/term/class and relation) and their version,

- Minimum machine-readable metadata to describe the semantic artefacts themselves and their content,

- Usage of **repositories/catalogs** to share, publish and retrieve semantic artefacts and their content

- Defining common API(s) to access and index semantic artefacts and their content,

- Interoperability approaches to make sure that semantic artefacts of various degrees of complexity and encoding format should work together including **publishing FAIR mappings and crosswalks** between semantic artefacts,

- Semantic artefacts and their content should be retrievable through search engines.

Related ongoing work

FAIR Impact

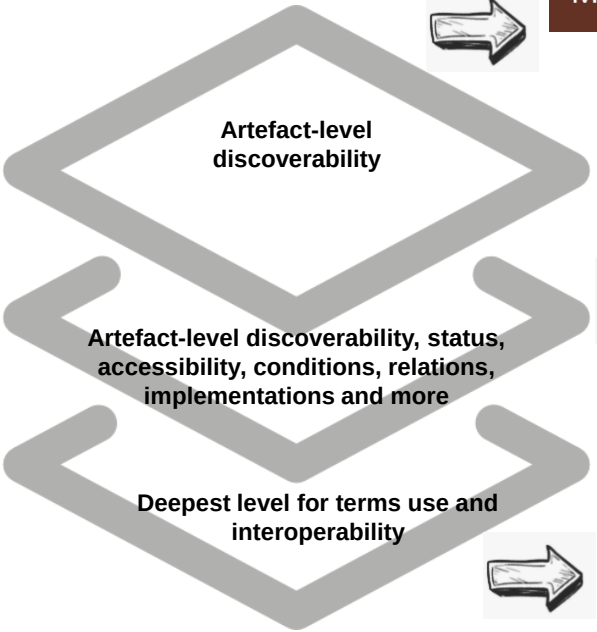
- M5.3 Semantic artefact assessment methodology: https://zenodo.org/record/8305173
- Common metadata schema for Semantic Artefact
- Semantic artefact governance
- FAIR Mappings
- Semantic artefact catalogs/repositories

FAIRCORE4EOSC - MSCR

Catalogues of semantic artefacts and their governance

Presented by Susanna-Assunta Sansone

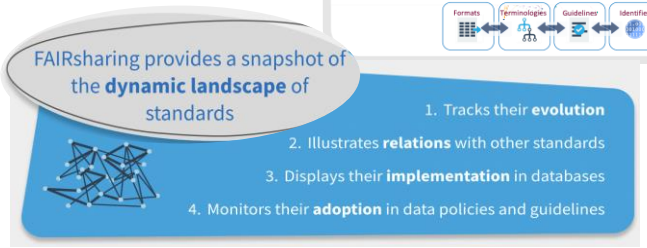
Types of catalogues/registries - examples



Metadata Standards Catalog

https://rdamsc.bath.ac.uk/

FAIRsharing.org



https://fairsharing.org/

BioPortal

Ontologies Search Annotator Recommender Mappings

https://bioportal.bioontology.org/

Sustainability aspects

Content

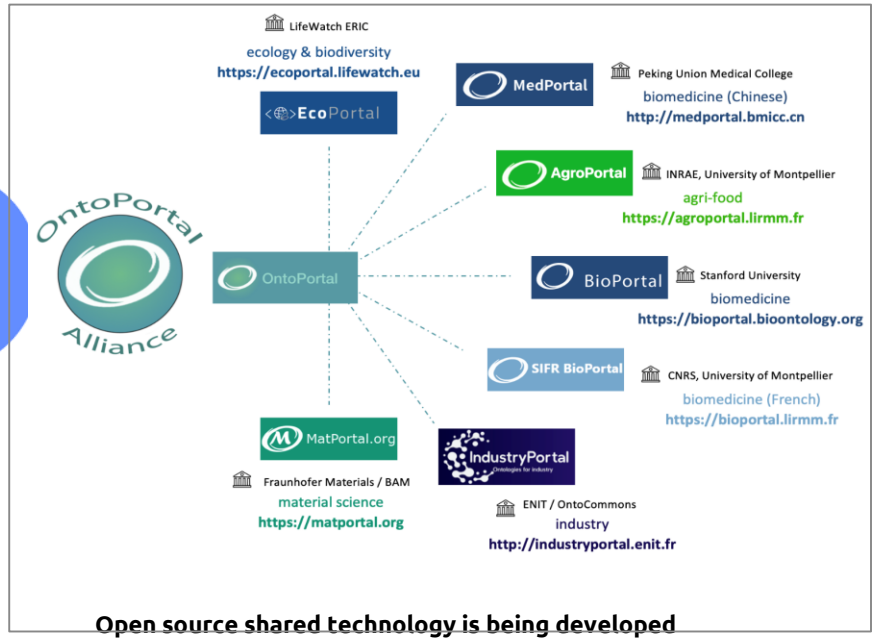
Technical

Embedding

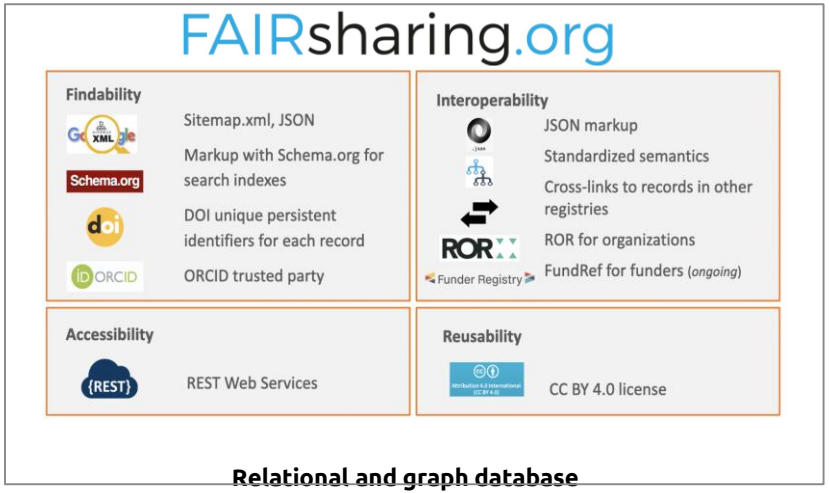
Endorsements

People and funds

Software diversity – examples



Open source shared technology is being developed



Relational and graph database

Metadata Standards Catalog

Github, API

Governance challenges - CLARIN example

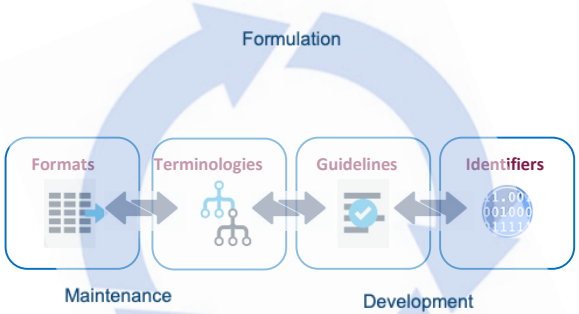
- **CLARIN's goal: make metadata profiles and language data explicit:**
  - Collaboration with ISO TC37 to try to use the ISOcat Data Category Registry[1]
    - too complicated/formalised ISO procedures to add/change concepts
    - categories (semantic artefacts/data model) not the concepts CLARIN needed
    - work too demanding, but useful in terms of what costs/efforts are involved
    - culture clash between the research community's dynamic needs and ISO's slow pace
  - Creation of the CLARIN Concept Registry[2], pragmatic and experts-driven
    - a pragmatic ontology, built with linguistic expertise from a editorial board
    - national content coordinators and forum; dealing with ISOcat legacy terms
    - guidelines and approval process for terms (add/change/expire)
    - community involvement with discussion and votes, although sustainability is challenging
- **Lesson learned => technology is usually not the main problem!**

1. http://www.lrec-conf.org/proceedings/lrec2014/pdf/153_Paper.pdf
2. https://pure.knaw.nl/ws/portalfiles/portal/1686735/CAC_2015_CCR.pdf

Organizational differences – SDOs vs grass-roots

Standard organizations, e.g.:

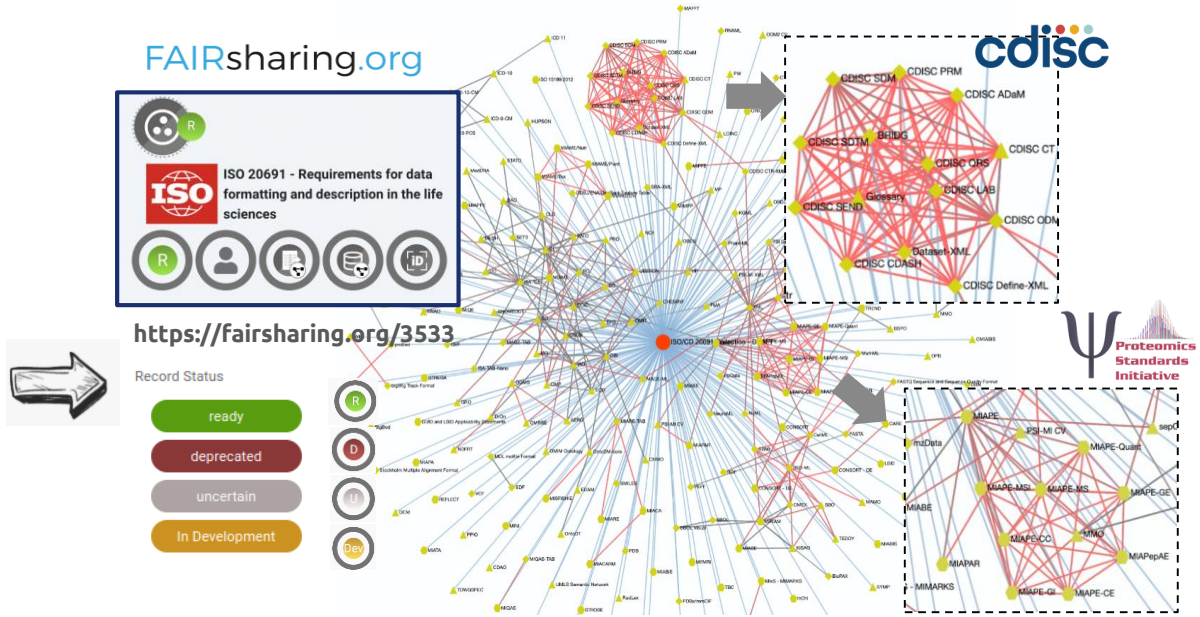
Grass-roots groups, e.g.:



- Industry-level standards
- Mostly regulators-driven
- Participation is often regulated
- Standards are sold or licenced
- Formal development process, often less flexible, could be lengthy
- Charges apply to advanced training or programmatic access

- Mostly research-level standards
- Open to any interested party
- Volunteering efforts
- Standards are free for use
- Development process varies, more flexible and adaptable to changes
- Minimal or little funds for carry out the work, let alone provide training

Synergies opportunities - ISO & FAIRsharing example



- FAIRsharing holds a 'live' collection of the standards (incl. semantics artefacts) and their relationships, recommended by the ISO document.
- Complements the static (and non-machine readable) list of standards in the ISO spec!

Long standing issues and known pain-points

Technical and *social* challenges, incl.:

- **Governance** and ownership, especially when working across-sectors
- Fragmentation, harmonization and extensions
- Indicators and evaluation methods
- Implementations, tools and services
- Credit and incentives for contributors
- Education, documentation and training
- Funding stream to support the ‘life cycle’ and uptake/adoption
- Business models for the sustainability of people/experts and related services

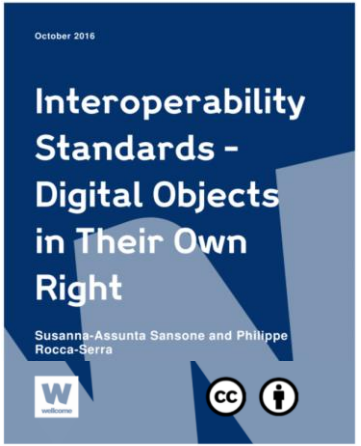
Empowering industrial research with shared biomedical vocabularies

Lee Harland[1,10], Christopher Larminie[2], Susanna-Assunta Sansone[3], Sorana Popa[4], M. Scott Marshall[5], Michael Braxenthaler[6], Michael Cantor[7], Wendy Filsell[8], Mark J. Forster[9], Enoch Huang[10], Andreas Matern[11], Mark Musen[12], Jasmin Saric[13], Ted Slater[14], Jabe Wilson[15], Nick Lynch[16], John Wise[17] and Ian Dix[18]

[1] Connected Discovery Ltd., 27 Old Gloucester Street, London WC1N 3AX, UK
[2] GlaxoSmithKline, Computational Biology, 2F157 Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK
[3] Standards and Data Sharing Infrastructure Team, e-Research Centre, University of Oxford, 7 Keble Rd, Oxford OX1 3QG, UK
[4] Knowledge Management and Information Science, R&D Information, AstraZeneca R&D Mölndal, 431 83 Mölndal, Sweden
[5] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands
[6] Pharma Research and Early Development, Hoffmann-LaRoche Inc., 340 Kingsland St, Nutley, NJ 07110, USA
[7] Pfizer Worldwide Research and Development, 235 E 42nd ST, MS 150/5/60N, New York, NY 10017, USA
[8] Unilever R&D, Colworth Science Park, Sharnbrook, Bedfordshire MK44 1LQ, UK
[9] Syngenta R&D Information Systems, International Research Centre, Jealott's Hill, Berkshire RG42 6EX, UK
[10] Pfizer Worldwide Research and Development, 35 Cambridge Park Drive, Cambridge, MA 02140, USA
[11] Thomson Reuters Life Sciences, 22 Thomson Place, Boston, MA 02210, USA
[12] Stanford University, Stanford University, 251 Campus Drive, Stanford, CA 94305-5479, USA
[13] Scientific Information Centre, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany
[14] Merck Sharp & Dohme Corp., 33 Avenue Louis Pasteur, Boston, MA 02115-5727, USA
[15] Science & Technology, Corporate Markets, Elsevier Pharma and Biotech Group, Elsevier, 32 Jamestown Road, London NW1 7BY, UK
[16] AstraZeneca UK, Alderley Park, Macclesfield SK10 4TG, UK
[17] The Pistoia Alliance[3]

Drug Discovery Today • Volume 16, Numbers 21/22 • November 2011

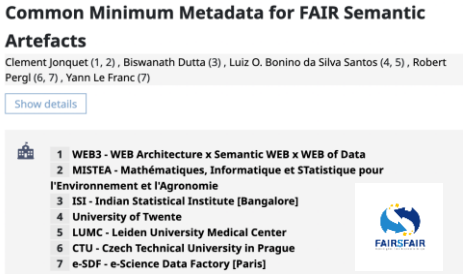
DOI: 10.1016/j.drudis.2011.09.013



DOI: 10.6084/m9.figshare.4055496.v1

Ongoing and planned activities – examples

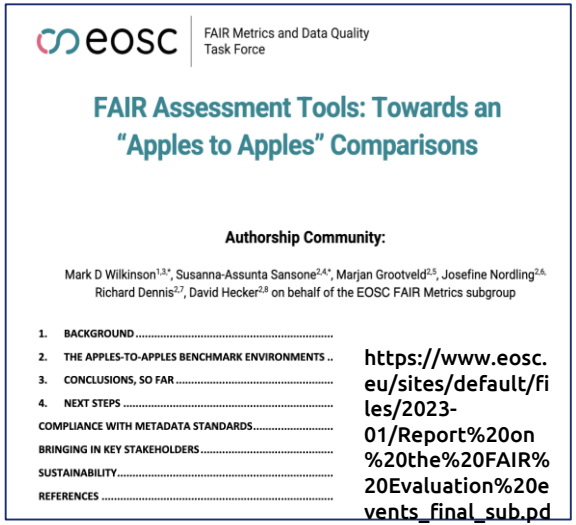
https://fair-impact.eu/events/fair-impact-events/fair-impact-semantic-artefact-governance-workshop`



https://hal.science/hal-04106533v2

Standards and Semantic artefacts are pillars of FAIR and therefore addressed by all EOSC projects

Semantic artefacts key in the FAIR evaluation/assessment



https://www.eosc.eu/sites/default/files/2023-01/Report%20on%20the%20FAIR%20Evaluation%20events_final_sub.pd

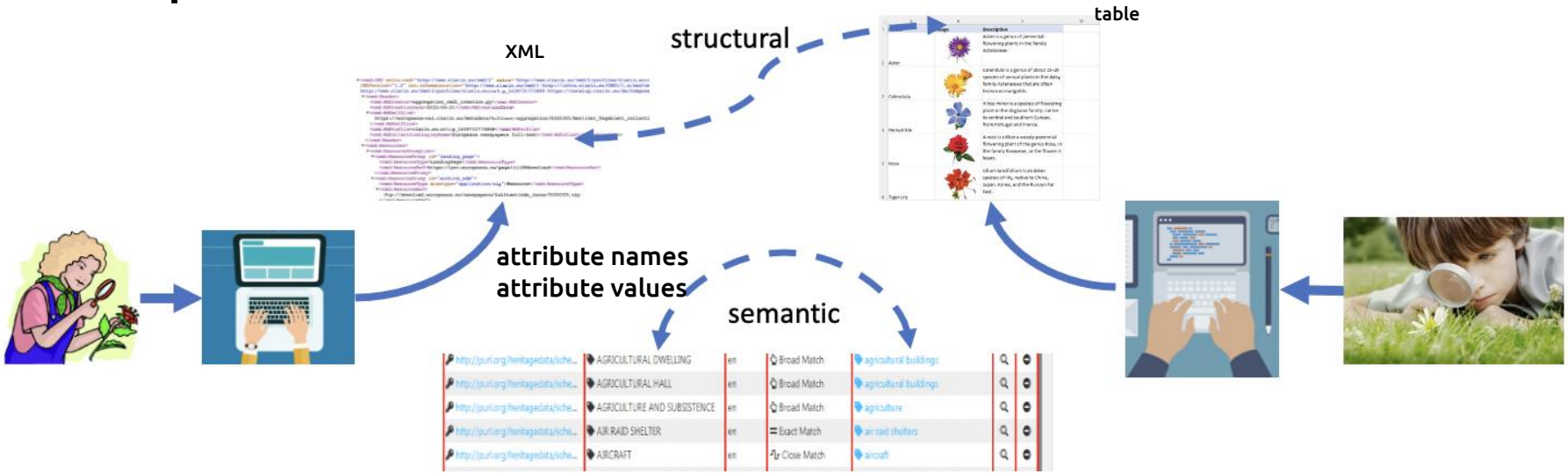
FAIRsharing.org

Mappings, crosswalks and alignment

Presented by Daan Broeder

Mapping different structures and semantics

Mappings relate the content of items in different data description schema meant for similar information.



semantic relation types: “equality (close) (exact)”, “broader” and “narrower” ...

value scheme conversions: eg. date/time, coordinate system formats
unit conversions: eg. temperature in K vs C

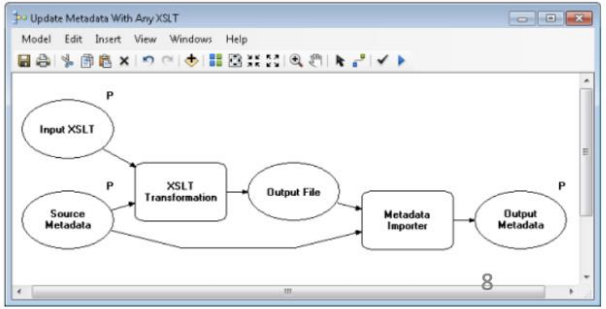
Mapping examples

Entity 1	Entity 2
tectonic movement(ENVO:01001093)	Continental drift (SWEETPhenGeolTectonic:ContinentalDrift)
river bank (ENVO:00000143)	Riparian zone (SWEETRealmLandCoastal:RiparianZone)
marine benthic biome (ENVO:01000024)	Benthic zone (SWEETRealmOcean:BenthicZone)
leaf alternate placement(FLOPO:0001032	Phyllotaxy (TO:0006014)
rhizome mass (FLOPO:0003190)	Rhizome dry weight (TO:0000556)
whole plant lifestyle (FLOPO:0980070)	Life cycle habit (TO:0002725)

simple mapping example BioDiv / Earth System Sciences

tables are powerful and can provide sufficient information (adding extra columns) for very detailed mapping descriptions

Often mapping specifications and conversion are combined in technologies such as XSLT



Some communities use XML as a metadata format with hierarchical schema, which can introduce additional context for the entity semantics and their mappings

not easy to create conversions without using complex code

The use for mappings and crosswalks

Main purposes:

- aggregating and integrating data-sets for further processing
- retrieval of suitable records from heterogeneous sources

Crosswalks do not always need to be complete, just need to achieve their (project limited) purpose, many are ‘experimental’

Note the difference between modelling and only specifying relations, and actually converting data, collecting all relevant mappings (crosswalk) and putting this in code

Mapping representations

Mappings can be represented or found in many formats

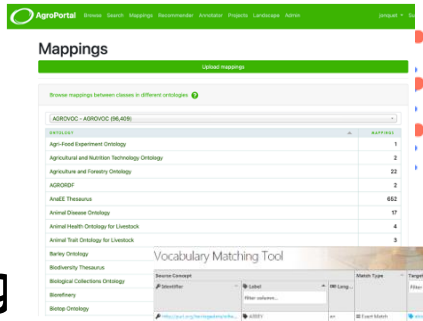
- Research papers, Tables, Code and XSLT, RDF/OWL
- Tables are still broadly mappings
- Source code and esp. XSLT also are heavily used

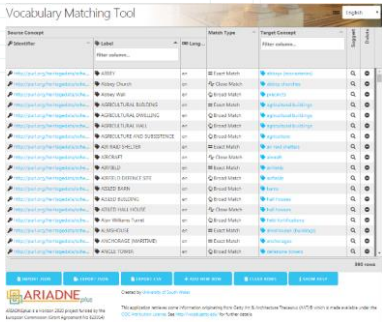
Mappings in non-RDF formats should not be ignored, improve their FAIRness by providing proper metadata, provenance and registration using PIDs

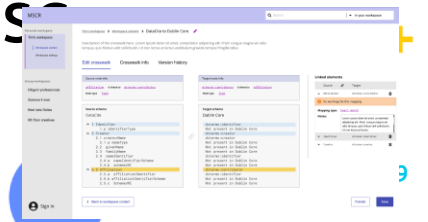
However proper modelling with agreed relation typology should be preferred for future work eg. using SSSOM

Tools for creating mappings and crosswalks

- Automatic, usually based on automatic string matching of the entities in semantic artefacts.
  - e.g. OntoPortal; challenges are the string matching to be domain specific fuzzy, and the lack of context
- Tools allowing experts to create relations between suitably visualised entities in schema & ontologies
- Good examples exist eg. Cocoda (libraries), VMT (Archeology/Cultural Heritage), DME (Humanities),
- New tool under construction, FAIRCORE4EOSC project is working on MSCR







Current relevant EOSC project work

MSCR
EOSC Metadata Schema & Crosswalk Registry

FAIRIMPACT:

See the references doc

FAIRCORE4EOSC:

- Creating the MSCR, a metadata schema crosswalk registry for all types of mappings including those embedded in texts, tables, code, …
  - FAIR data management functions, ao: proper metadata, provenance and PIDs helping researchers to find, reuse and share crosswalks
  - A graphical mapping tool allowing users create crosswalks between different metadata schema
- Two community case-studies Climate and SSH testing out the MSCR
- Two demonstrators, general data management tools B2SHARE and B2FIND using the MSCR to manage metadata schema and mappings

CLARIN

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

B2FIND

B2SHARE

eosc | FAIR-EASE

National Oceanography Centre

British Oceanographic Data Centre

Implementation examples: The Semantic Analyser

Alexandra Kokkinaki & Gwenaelle Moncoiffe

eosc

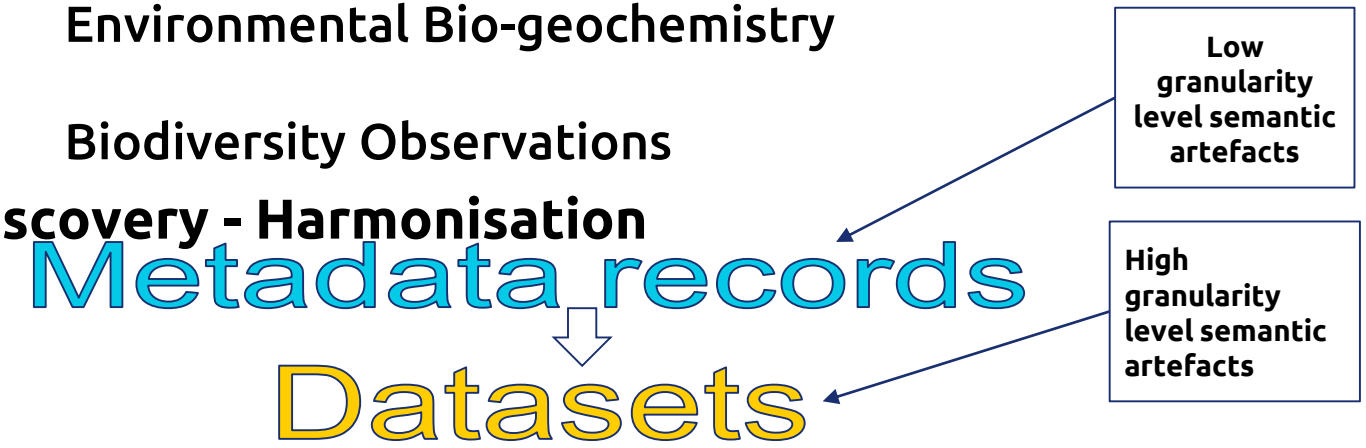
EOSC Future

EU23 SPANISH PRESIDENCY COUNCIL OF THE EUROPEAN UNION

Funded by the European Union

The objective

- **Semantic interoperability** for multidisciplinary data from various sources
  - Earth & Environmental Dynamics
  - Environmental Bio-geochemistry
  - Biodiversity Observations
- **Discovery - Harmonisation**



Building Blocks

- **Crosswalks** between metadata/data standards
  - e.g. from ISO19139, DCAT etc TO ISO19115
- **Mappings** between semantic artefacts (terms/instances)
- Avoid manual mappings when possible (tedious)
  - Mapping frameworks e.g. I-ADOPT

Syntactic

Semantic

Mappings

Crosswalks

Values

Strings that may/may not originate from a semantic artefact

oceans, geoscientificInformation,sea surface temperature,fCO2,Ship,satellite-observation,

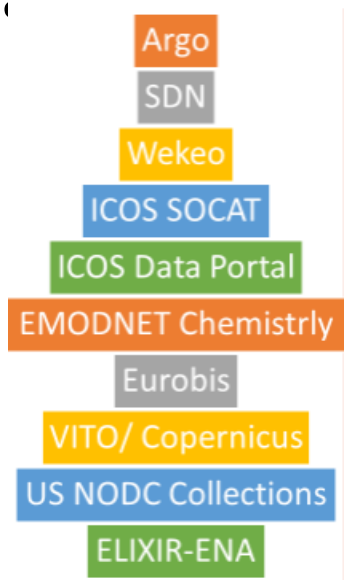
Seas and coasts etc,marine biome (ENVO:00000447), 454 GS FLX Titanium,

Inorganic chemical composition of sediment or rocks

URIs: http://vocab.nerc.ac.uk/collection/L05/current/60/ http://www.seadatanet.org/urnurl/SDN:P02::RMIN/

Codes: ENVO:00000447, SDN\:L05\:\:60

Combination: deep chlorophyll maximum layer (ENVO:xxxxxxxx)



The semantic analyser:

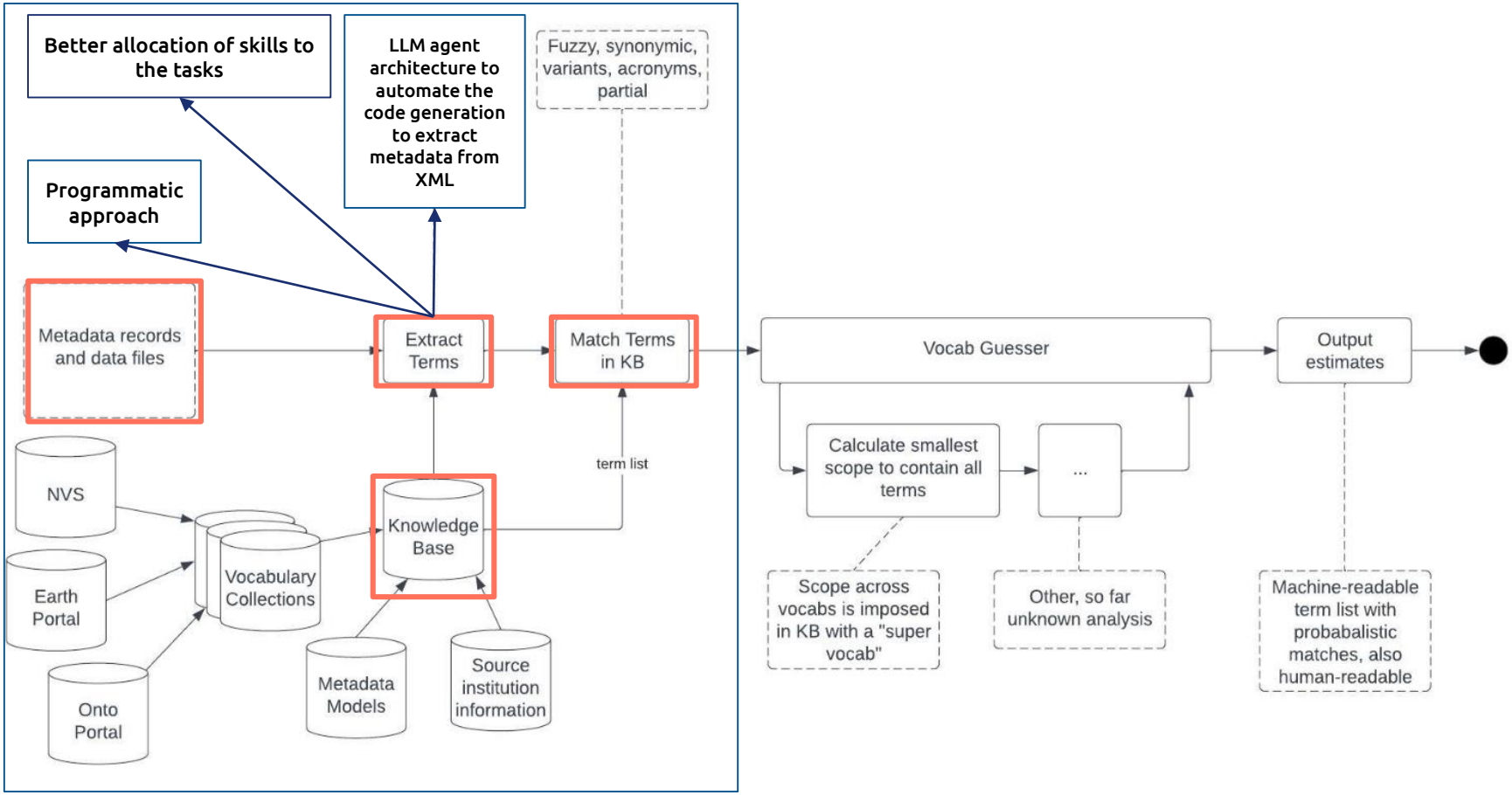
Focus on a set of metadata elements

- Keywords
- Parameters
- Units
- Platforms
- Sensors

Analyse them to identify

Semantic artefacts?

```
<gmd:descriptiveKeywords>
  <gmd:MD_Keywords>
    <gmd:keyword>
      <gco:CharacterString>Biocenosis</gco:Ch
      aracterString>
    </gmd:keyword>
    <gmd:keyword>
      <gco:CharacterString>Eutrophication</
      gco:CharacterString>
    </gmd:keyword>
    <gmd:keyword>
      <gco:CharacterString>EUROBIS</gco:Chara
      cterString>
    </gmd:keyword>
  </gmd:MD_Keywords>
</gmd:descriptiveKeywords>
```


Better allocation of skills to the tasks
LLM agent architecture to automate the code generation to extract metadata from XML
Fuzzy, synonymic, variants, acronyms, partial
Programmatic approach
Metadata records and data files
Extract Terms
Match Terms in KB
Vocab Guesser
Output estimates
Calculate smallest scope to contain all terms
...
term list
NVS
Knowledge Base
Earth Portal
Vocabulary Collections
Onto Portal
Metadata Models
Source institution information
Scope across vocabs is imposed in KB with a "super vocab"
Other, so far unknown analysis
Machine-readable term list with probabalistic matches, also human-readable

British Oceanographic Data Centre

Sources

SELECT ALL

- [ ] ELIXIR-ENA
- [x] ICOS SOCAT
- [ ] ICOS Data Portal
- [ ] Copernicus Marine Environment Monitoring Service (CMEMS)
- [x] SeaDataNet products
- [ ] EMODnet Chemistry

SEARCH

Results

PREVIOUS PAGE NEXT PAGE

8E494D045ED3115EEF5C6E697510408A52DA203

SOCAT v2020 DSG Files

BBC7EF49B0FA24B73C286BB58999F8FC1B34DC

XUE08_064-104

SOCAT v2020 DSG Files

41C44BA3E831654BCCA93CFF1ADA2FC17D0CED

SOCAT v2020 DSG Files

sdn-open:urn:SDN:CDI:LOCAL:5060-5060-5060-ds0

SeaDataNet - Marine geology from Institute for Ma

Analyser

Analyser Endpoint

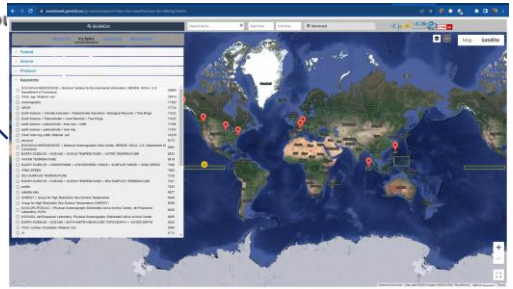
Enter URI

Threshold

1

ANALYSE METADATA

Analyser o



The semantic analyzer

British Oceanographic Data Centre

Sources

SELECT ALL

- [ ] ELIXIR-ENA
- [ ] ICOS SOCAT
- [ ] ICOS Data Portal
- [ ] Copernicus Marine Environment Monitoring Service (CMEMS)
- [x] SeaDataNet products
- [ ] EMODnet Chemistry

SEARCH

Results (973)

sdn-open:um:SDN:CDI:LOCAL:1609-1609-1609-ds04-4
SeaDataNet - Marine geology from Geological Survey
SeaDataNet is the Pan-European infrastructure for marine and ocean data management and delivery serv...

sdn-open:um:SDN:CDI:LOCAL:1022-1426-1022-ds04-4
SeaDataNet - Marine geology from French Geological
SeaDataNet is the Pan-European infrastructure for marine and ocean data management and delivery serv...

sdn-open:um:SDN:CDI:LOCAL:1022-485-1022-ds12-4
SeaDataNet - Human activities from French Geologica
SeaDataNet is the Pan-European infrastructure for marine and ocean data management and delivery serv...

sdn-open:um:SDN:CDI:LOCAL:1022-1926-1022-ds08-4
SeaDataNet - Terrestrial from French Geological Surv
SeaDataNet is the Pan-European infrastructure for marine and ocean data management and delivery serv...

sdn-open:um:SDN:CDI:LOCAL:96-1570-96-ds03-4
SeaDataNet - Physical oceanography from German O
SeaDataNet is the Pan-European infrastructure for marine and ocean data management and delivery serv...

sdn-open:um:SDN:CDI:LOCAL:120-4751-120-ds12-4
SeaDataNet - Human activities from National Institute

1 files selected:
- SeaDataNet - Biological oceanography from CNR, Natio... *(83237 bytes)*

Analyser

Analyser Endpoint: https://99koor0nmj.execute-api.ap-southeast-2.amazonaws.com/production/process_metadata

Threshold: 1

ANALYSE METADATA

Analyser output

SeaDataNet - Biological oceanography from CNR, National Research Council, Institute of Marine Science (Ancona) (PointOfContact; Data Custodian; Data Distributor), point observations

Method	MatchURI	MatchProperty	MatchTerm	SearchTerm
Instrument - Exact Match	http://vocab.nerc.ac.uk/collection/L05/current/351/	preferred label	SDN:L05::351	SDN:L05\:\:351
Instrument - Exact Match	http://vocab.nerc.ac.uk/collection/L05/current/30/	preferred label	SDN:L05::30	SDN:L05\:\:30
Instrument - Exact Match	http://vocab.nerc.ac.uk/collection/L05/current/113/	preferred label	SDN:L05::113	SDN:L05\:\:113
Instrument - Exact Match	http://vocab.nerc.ac.uk/collection/L05/current/23/	preferred label	SDN:L05::23	SDN:L05\:\:23
Instrument - Exact Match	http://vocab.nerc.ac.uk/collection/L05/current/90/	preferred label	SDN:L05::90	SDN:L05\:\:90
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/130/	preferred label	SDN:L05::130	SDN:L05\:\:130
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/351/	preferred label	SDN:L05::351	SDN:L05\:\:351
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/30/	preferred label	SDN:L05::30	SDN:L05\:\:30
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/113/	preferred label	SDN:L05::113	SDN:L05\:\:113
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/23/	preferred label	SDN:L05::23	SDN:L05\:\:23
Instrument - Exact Match (Identifiers)	http://vocab.nerc.ac.uk/collection/L05/current/90/	preferred label	SDN:L05::90	SDN:L05\:\:90
Variable - Exact Match	http://vocab.nerc.ac.uk/collection/P03/current/D032/	preferred label	Sea level	sea level
Variable - Exact Match	https://gcmd.earthdata.nasa.gov/kms/concept/9ac7a1c5-4179-47bc-8589-ebaa90d6cbd1	preferred label	SEA LEVEL	sea level

The Knowledge Base

https://w3id.org/ozcar-theia
http://inspire.ec.europa.eu/theme
https://w3id.org/ozcar-theia/
http://purl.obolibrary.org/obo/envo.owl
http://www.w3.org/ns/sosa/
http://www.w3.org/ns/sosa/
http://www.w3.org/ns/ssn/
http://www.w3.org/ns/ssn/
http://qudt.org/2.1/vocab/unit
http://www.w3.org/1999/02/22-rdf-syntax-ns#
http://www.w3.org/2000/01/rdf-schema#
http://www.w3.org/ns/prov-o#
http://www.w3.org/2002/07/owl
http://www.w3.org/2004/02/skos/core
http://purl.org/voc/cpm
https://w3id.org/iadopt/ont
https://w3id.org/iadopt/ont
http://www.w3.org/ns/prov#
http://qudt.org/2.1/vocab/quantitykind
http://www.w3.org/ns/dcat
https://w3id.org/ozcar-theia/variableCategoriesGroup
https://w3id.org/ozcar-theia/variableGroup
https://w3id.org/ozcar-theia/skosCollection_a415bff4
https://w3id.org/ozcar-theia/skosCollection_58254f70
https://w3id.org/ozcar-theia/skosCollection_65cf7fa2
https://w3id.org/ozcar-theia/skosCollection_7198f400
https://w3id.org/ozcar-theia/skosCollection_ba31712c
http://vocab.nerc.ac.uk/collection/

Next steps for the semantic analyser (SA)

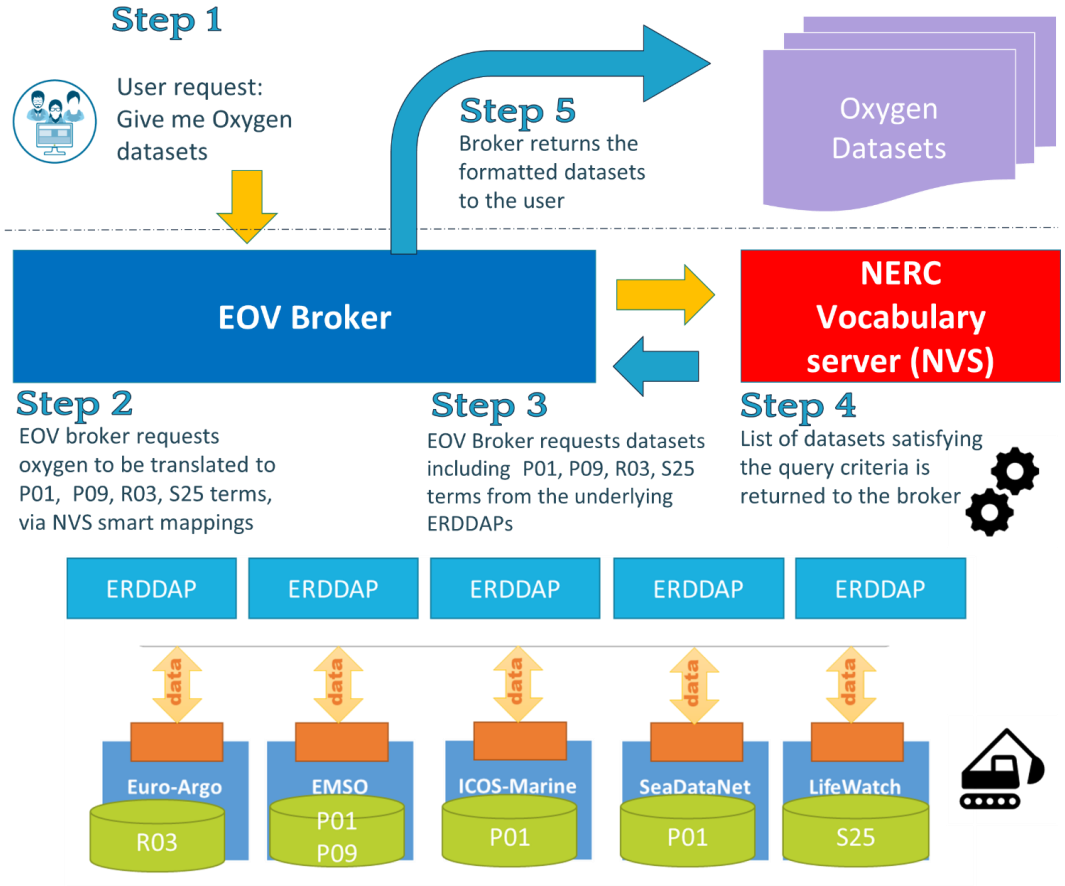
- Enhance the KB with more semantic resources as we learn more about the data
- Finalise the LLM algorithm to automatically identify diverse XML snippets
- Enhance the SA input with datasets
- Implement the vocab guesser to provide mapping suggestions

The EOV demonstrator - parameter harmonisation

i-adopt

Smart mappings

Find all the datasets that observe what is defined to be EOV oxygen



Conclusions

Complying to a standard does not achieve interoperability:

- Community agreements are essential

Semantic resources need standardised referencing from metadata/data records

- see definedTerm

Semantic mappings are a way to provide harmonisation but they are tedious:

Frameworks like **iadopt** can provide:

- Automated mappings
- Sliding up and down the granularity scale

#EOSCsymposium23

Next up

EOSC contribution to Research Assessment

Barcelona | 12.10 - 13.10

eosc

EOSC Future

EU23 SPANISH PRESIDENCY COUNCIL OF THE EUROPEAN UNION

Funded by the European Union

Stay tuned

symposium23.eoscfuture.eu