eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

1st Synchronisation Force Workshop 2022

Session 4 on Metadata, semantics and interoperability.

Thursday 24th November, 14.00-15.30 CET

Clement Jonquet (INRAE) &
Oscar Corcho (UPM)

Funded by the European Union

Some details for this session.

- The session will be recorded but only for internal use for the rapporteur and the report writing.
- Shared **spreadsheet** and **note taking document** for use in the session (and until November 28th (CoB)).
  - Survey responses have been added to the spreadsheet.
  - Keep information factual, short and include links wherever possible.
  - In the note taking document you may add more detail and background.
  - Be careful when editing spreadsheet cells.
  - Refrain from editing other people's information.

Session questions

1. **Which are the semantic artefacts** that you develop or use in your scientific domain and how are they **governed** (developed and maintained)?

2. What **limitations** do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

3. Is there any type of **semantic artefact catalogue** (vocabulary or terminology service, ontology library or repositories, etc.) in your scientific domain, where some (or all) of these semantic artefacts are catalogued?
    a. What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

4. Is there a strategy to deal with **crosswalk and mappings between semantic artefacts** in your scientific domain?

A couple of definitions (not absolute, but during this session)

Semantic artefacts: a broader term to include ontologies, terminologies, taxonomies, thesauri, vocabularies, metadata schemas and standards.

Legacy of FAIRsFAIR and adopted in the EOSC Interoperabibily Framework

Semantic artefact catalogues: encompass any existing ontology repositories, registries, vocabulary/terminology services and metadata schemas catalogues.

(Semantic) Crosswalks and mappings: formal links between the content of these semantic artefacts.

FAIRsFAIR D2.5 FAIR Semantics Recommendations

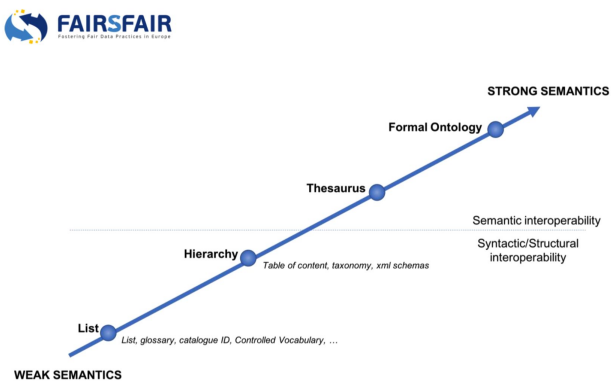


Figure 2: Semantic artefact spectrum. Derived from Leo Obrst, 2010

A semantic artefact is defined in this 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, e.g., RDF Turtle, and OWL, using XML (RDF) and JSON-LD.

Session featured panelists

- 8’ Biomedicine (Nicolas Matenzoglu & Pier Luigi Buttigieg)
- 8’ Ecology/biodiversity (Naouel Karam & Ilaria Rosati)
- 8’ Agri-food (Clement Jonquet)
- 8’ Social sciences & humanities (Arnaud Gingold)
- 8’ Industry (Hedi Karray)
- 8’ Astronomy (Baptiste Cecconi)
- 8’ Earth Sciences (J-C Desconnets, V. Agazzi, C. Pierkot)

Biomedicine

(with bridges to, food science, Earth & environment / UN Sustainable Development etc etc)

(Nicolas Matentzoglu & Pier Luigi Buttigieg)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

- We are members of the OBO Operations Committee, which itself organises a registry of semantic artefacts for the life sciences, with bridges to other domains (e.g. UN Data Systems, Earth and Environment, Biodiversity, Sustainable Development)
- OBO Foundry itself is not involved with governance processes of specific ontologies, but the ones Pier and Nico are involved in (e.g. Uberon, ENVO, Mondo, CL, uPheno, PATO, etc) tend to be community-driven (community members open PRs)
    - Uberon for example at least a dozen large stakeholder organisations, and the central management is done through a coordination call
    - ENVO is deployed in biomedical and environmental encourages pull requests and review by an editorial team; we're refining that model to cope with more diverse and rapid input from our userbase and CoPs
- Most OBO projects use ROBOT for defining the ontology life-cycle processes, and many use ODK for standardised life-cycle workflows.
- Most OBO ontologies are build using standard GitHub collaborative workflows.

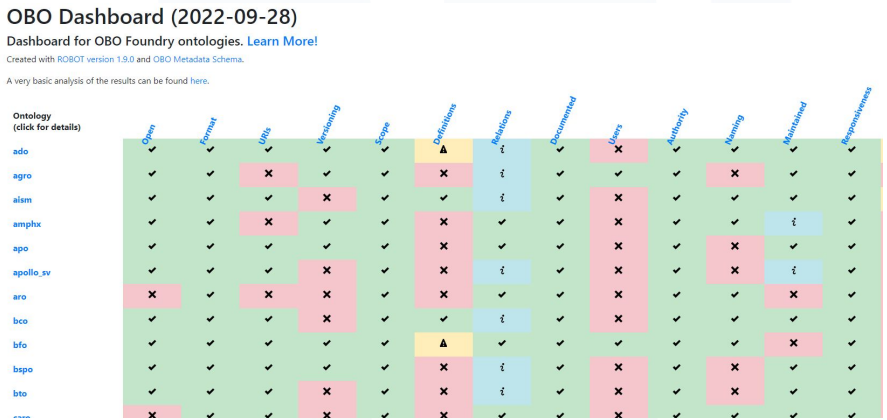
What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

- Very little validation of semantic metadata, i.e. domain/range restrictions in slots of semantic data models (GWAS “trait” example)
- Poor quality definitions (human and machine readable) resulting in odd annotations
- Reluctance to co-develop existing solutions, rather than create (often rushed and unmaintained) new resources
- Yawning gap between quantitative data (measurements) and qualitative data (phenotype/disease)
- Many overlapping ontologies/vocabularies with poor mappings
  - confusion in generalised curation tools
- Despite serious efforts, metadata in and about semantic artefacts are still too variable for generalised tooling, e.g.
  - Obsoletion patterns (automatic replacement pipelines)
  - Attribution

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

- http://obofoundry.org/
- Carefully curated catalog of (primarily) biomedical ontologies

With consistent, automated reporting on each ontology's status (vs the OBO Principles) in a dashboard: http://dashboard.obofoundry.org/dashboard/index.html



What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

- OBO Foundry Registry is doing quite well as a semantic artefacts catalogue (purl system, detailed metadata, website)
- Lack of incentive for ongoing improvement and metadata convergence (high cost!) across catalog entries (ontology and term level metadata)
- Semantic artefacts catalogues like OBO Foundry:
  - still struggle with a universal quality indicator (how mature/good is this ontology? Not just metadata, but content)
  - are not great at communicating overlap across artefacts (which disease ontology should I use?)
  - tend to be siloed, i.e. not well aligned with other catalogues
  - should also capture ontology design patterns but they dont, which leads to incompatible axiomatisation

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

- Term “mapping” ill defined - entity mappings, alignment, schema mapping entirely different
- We recently started promoting a new FAIR standard for sharing semantic mappings: https://github.com/mapping-commons/sssom
  - Used by other communities, such as the official MIxS DwC mapping (biodiversity): https://www.tdwg.org/community/gbwg/MIxS/
- A tool ecosystem is emerging around this:
  - matching (OAEI, OAK)
  - transformation/validation (sssom toolkit)
  - maintenance and dissemination (mapping commons)
  - reconciliation (boomer, sssom toolkit)
- However, most links across ontologies in our domain are mostly of “logical” nature (“amount” is a characteristic used to quantify a “chemical” which accumulates in an “organ tissue”), because of the orthogonality principle - for this, the answer would be “shared design patterns” rather than “mappings”
- Generally, usage workflows for mappings are ill-defined (KG merging, data transformation, annotation grouping, etc)

Ecology/biodiversity

(Naouel Karam & Ilaria Rosati)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

- Thesauri on functional traits of aquatic organisms and on alien species developed for LifeWatch Italy using the editing tool VocBench and published through the repository of semantic resources for the ecological domain EcoPortal
- EnvThes, BODC vocabularies, DwC controlled vocabularies and others available through Ecoportal or external semantic services used to label data and in metadata
- CIDOC CRM, CRM digital, CRM science, Parthenos and LUPO ontologies used in the semantic model that describes different resources and their metadata in the LifeWatch Italy Semantic Platform
- Semantic artefacts developed in-house (e.g. LW Italy thesauri) follow the EcoPortal governance.
- All others follow guidelines and/or governance of a specific organization or initiative.

What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

- Difficulty to select the best term/concept due to a large collection of independent semantic artefacts in the same domain
- Too many semantic artefacts with the same terms/concepts without alignment
- Lack of semantic annotation tools/services and their integration on the data management systems
- Agreed strategy and governance needed

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

EcoPortal, BioPortal,AgroPortal, OBO Foundry, Gfbio Terminology Service, Research Vocabularies Australia, NERC Vocabulary Server, FAIRsharing.

What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

There is not a strategy for crosswalk and mappings in the scientific domain. We need to analysis what exists and adopt a common solution that could be integrated in our catalogues and tools.

At the moment, EcoPortal offers a basic mapping tool, but it should be improved in terms of algorithms, validation and other technical issues. Vocbench tool also allows the management of alignment from two different source: 1. File: loads a document expressed according to model of INRIA's Alignment API; 2. Remote Alignment System: allows to run an alignment task exploiting an external alignment system. This task, when completed, returns a list of alignments found and the project manager can validate or reject the proposed alignments one by one, or apply some quick actions (for more specifications see http://vocbench.uniroma2.it/doc/user/alignment_validation.jsf).

Agri-food

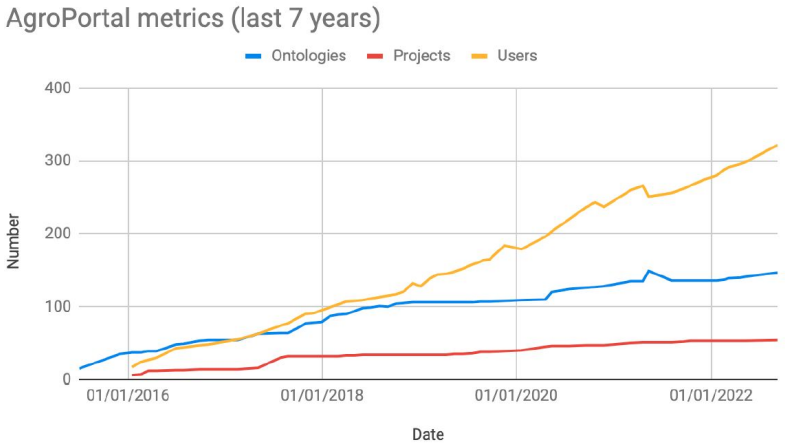
(Clement Jonquet)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

- Ontologies developed in with the OBO guidelines e.g., PO, TO, AGRO, FOODON
- Ontologies / trait dictionaries developed in group project such as Crop Ontology project
- Reference thesaurus developed by organizations: AGROVOC, INRAE Thesaurus, ANAEE Thesaurus
- Application ontologies e.g., OFPE, AFEO, PO2

Developed by different groups of scientists

- with specific guidelines and/or governance
- under the umbrella of a specific action or project
- often driven by concrete application tasks
- sometimes without any coordination
- no global governance



What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

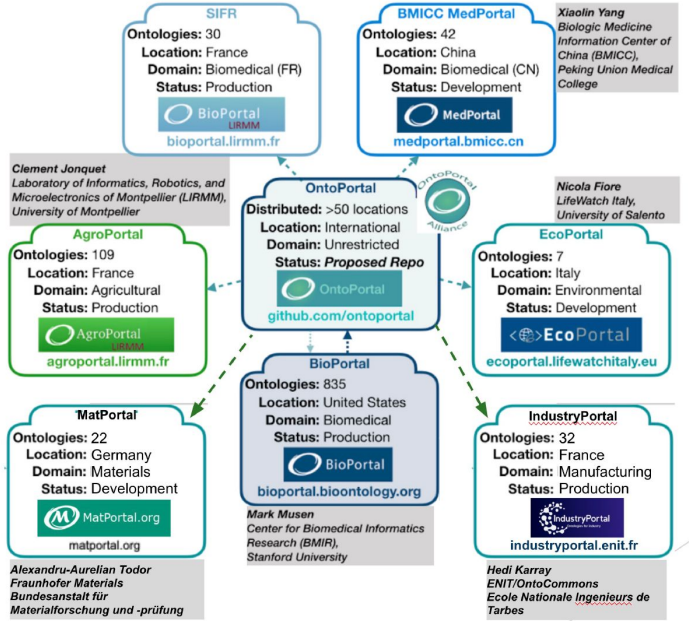
- Always changing contexts
- Long term availability, strategy and governance needed
- Problem of selecting Mr right semantic Artefact and Ms right ontologie
  - Then select the appropriate term/concept
- Need for a mapping repository
  - (beyond what's available for now in AgroPortal)
- Long term support of infrastructure to access and use SA
- Interoperation with data repositories is still hard

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

- With resources in the agri-food domain, we can cite:
  - SA libraries: OBO Foundry, FAIRsharing
  - SA repositories: EBI OLS, NCBO BioPortal, AgroPortal, etc..
  - specific web applications such as the Map of agri-food data standards, or cropontology.org

- AgroPortal develops a focus on agri-food with and open and use case driven approach
  - Also include a Recommender, an Annotator, a FAIRness assessment tool, a (simple) mapping repository

- Standardized approach with the OntoPortal Alliance effort. Towards an EOSC ready component?



What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

- Lack of coordination & governance
- Need for long term strategies and support. If SAs are really important research outputs, let's get organized for their support on the long term
- Still a variety of metadata approach and models (not restricted to agri-food)
- Taking into consideration the whole life cycle of SA (from designing to sharing)
- How to facilitate the recommendation and selection of the appropriate SA while taking the role of a scientific archive of what happens/happened in the domain?

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

- We need to all aspects related to mappings.
- SEMAF was an interesting study
-
- We will adopt SSSOM and recommend everyone to do so. Even if not perfect, that's the first real specification for exchanging mappings
  - FAIR mappings within FAIR-IMPACT?

- We need a strategy / governance that also covers mappings to avoid mapping the same SA again and again

- AgroPortal offers a mapping repository. A basic one, but at least mappings are reified into specific objects that can be identified and manipulated as any data.
  - Much improvement needed
  - Upload of SSSOM mappings in batch is now possible (but we loose some information as our model is not fully compliant)

Social sciences & humanities

(Arnaud Gingold)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

The project TRIPLE developed its own multilingual thesaurus for SSH contents (data and publications), based on LCSH thesaurus (multidisciplinary library thesaurus), with mappings. Set up by the consortium with manual review.

Available at:
https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries

In the SSH, semantic artefacts are often related to disciplines or domains (eg, linguistics, social sciences, archeology). Related to disciplines which traditionally use more and bigger datasets. Other SA are not scientific specific: coming from libraries, publishers, or about generic information (eg languages).

Indicative list: CLAVAS, CIDOC-CRM, Cessda Vocabulary Service, Pactols, Vocabs, DANS KOSo, Loterre, LCSH, RAMEAU

What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

In the SSH, the disciplinary boundaries make it difficult to have a unified environment: not only the perimeter but the granularity and the terminology changes. Having all existing SA connected is challenging.

Differences: data types, data collection, theories, methods

Also, a complete list of available resources in the SSH still has to be made.

Another challenge comes with multilingualism, which is in the SSH adds another level of complexity of the mappings (technical terms are more "universal").

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

Like said above: a complete list of available resources in the SSH still has to be made.

But catalogues exist: Cessda, CLARIN, Greek EKT, etc.

Depends probably also on the coordination of a network of SA experts in the SSH fields.

Another question would be, once identified these SA in SSH: who is using them and to which extent?

What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

That would probably be a project in itself.

In TRIPLE, we took a generic SA to cover all the SSH fields (LCSH thesaurus), along with mappings (French RAMEAU, Cessda ELSST).

LCSH lacks of specificity. Issues with terms too broad or common.

In other projects, the objective was to try to integrate many voc together, but issues with granularity and specific terms.

My guess would be to start in the middle: identify most used and general voc, find mappings and/or terms adaptations that could work for as much fields as possible.

Industry

(Hedi Karray)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

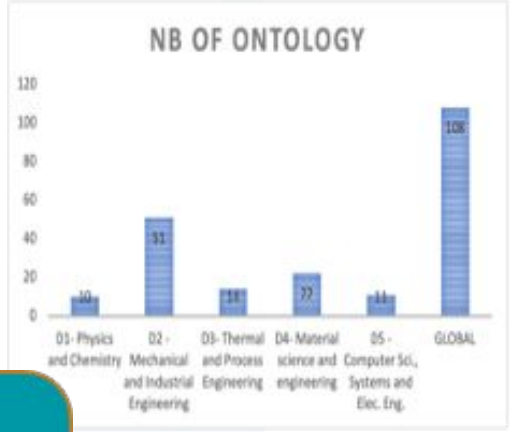
- Ontologies developed: ROMAIN, SIMPM, IOF-Core, Maintenance, Process Planning, Scheduling, product model, 3D environment, Health service, Crisis Management.

- Lots of other work has been done by community, but largely at a low Technology Readiness Level - TRL (i.e., research work), very little of this has direct economic commercial value.

- Uptake of semantic technology is still partial and there is a findability issue (see below)
- To improve the uptake of semantic technologies, their benefits need to be further explained/demonstrated to MM domain experts.

AIOTI, StandICT and OntoCommons

- OntoCommons landscape survey (D3.2 - 2022 https://zenodo.org/record/6504553)
  - 150 total ontologies, 90 machine readable in materials and manufacturing
  - Performed classification by topics, TLO usage, topology
  - FAIRness, coverage, overlap, semantic gaps, usage, maturity analysis
  - Sources stored in IndustryPortal and OntoCommons Registry.
- EUOS ontology survey (Stand-ICT) – ongoing, over 130 ontologies collected
- Gap Analysis - Semantic Interoperability in Practice https://tinyurl.com/799uy5kn
  - OntoCommons Roadmap (to be published)



Industrial Needs

- Data integration and sharing.
- Standardisation
- **Various domain Perspectives**
- **Interface domain ontologies with TLOs**
- Link domain experts to Ontologists

Gaps in Domain Ontology

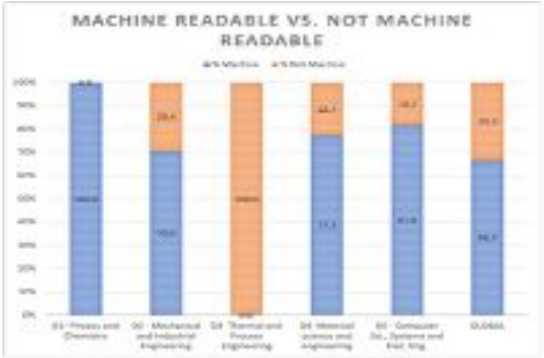
- Models granularity
- Lack of Generic and Application-specific Ontologies
- Lack of standardised methodology and tools
- Ontology as a conceptualization of reality vs information model
- Ontology Sustainability
- Lack of Standardised Method for Domain Ontology Evaluation

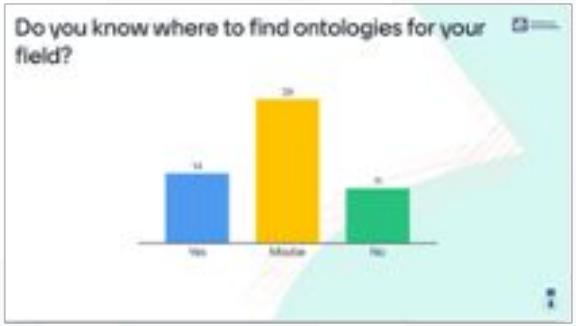
Recommended Actions

- Standardization of the ontology engineering steps
- TLO-MLO Alignment
- **Balance of Theory and Practice**
- **FAIRness**
- **Follow Domain related standards**
- **Classify domains**
- **Bridging the gap between domain experts and ontologists**

What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

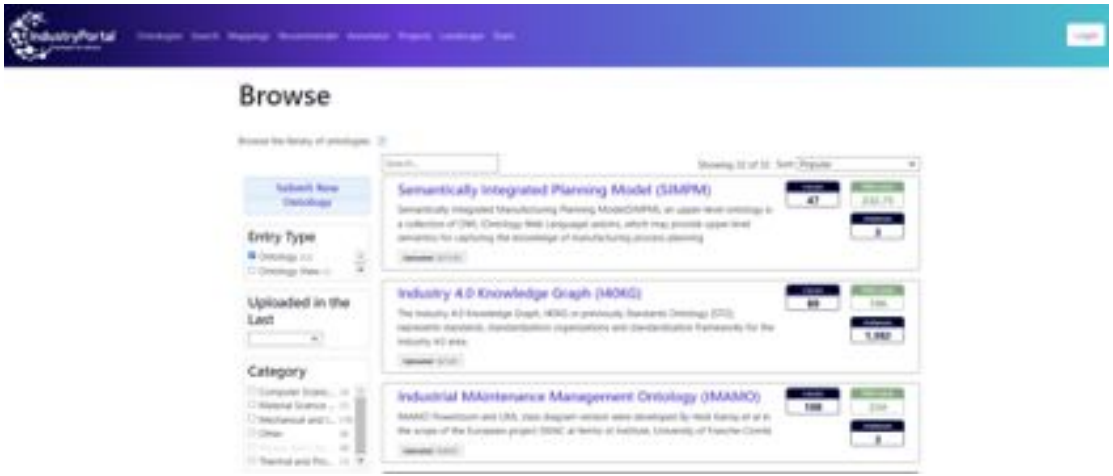
"one of the main services that the OntoCommons project can offer is to create a public resource where all the major ontologies, their uses, and their users, are listed. " – Barry Smith





Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

http://industryportal.enit.fr/



- Deployment: January 2022
- 52 Ontologies
- 10 Users
- Candidates: 50 Ontologies

-OntoCommons Registry :
https://data.ontocommons.linkeddata.es/index

IndustryPortal has adopted MOD1.4 driven FAIRness metadata and assessment tool O'FAIRe to ensure FAIRness

What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

- Such catalogue did not exist for industry before. If there are, they are either proprietary, have limited coverage, focuses on some specific domain(s), not maintained, and no update are provided.
- FAIRness is not addressed.
- Wrong reuse strategy (ontology reuse was predominantly interpreted as usage of arbitrary information sources of the relevant cases.)
- No commonly accepted quality control standards
- Poor training
- Poor documentation

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

- OntoCommons is developing a TRO (aligned set of TLOs) driven methodology and technical principle to harmonise ontologies in NMBP domain at different level of interoperability.
- IndustryPortal is building SSSOM driven metadata for mapping (OntoCommons specific) for alignment of ontologies at top, middle and domain levels, along with automatic dependency resolution.
- A federated toolchain is being built to provide end-to-end guided workflows for ontology development, harmonization and data documentation to help users perform tasks specific to ontology lifecycle phases using various tools and best practices recommended by OntoCommons, such as UPM tools (CHAWLK, THEMIS, OOPs, FOOPs, WIDOCO, ...), ENIT tools (IndustryPortal, SousLeSens), and templates (ORSD, Bridge Concept Template...).
- Methodology: consolidation of LOT (LOT4OntoCommons), OntoCommons Technical Principle, OntoCommons Bridge Concept Methodology.

Astronomy

(Baptiste Cecconi)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

Well defined interfaces, various protocols and standards ⇒ data and information models.
Same language = english for all

- **Controlled list of terms** in information models (roles, types, names…)
- **Thesaurus** (for subjects: e.g., Unified Astronomy Thesaurus)

Managed by international alliances (in working groups, decisions by voting or consensus):
IVOA (astrophysics); IPDA (Planetary Sciences), IHDEA (Heliophysics)

Examples

- IVOA vocabularies: lists of reference frames, time scales, subjects, product types, object types…
- SPASE Data model (IHDEA): lists of measurement types, locations and regions, spectral ranges…
- PDS4 Information model (IPDA): lists of observing system component types, spectral ranges, spatial domain…

Name resolvers:

- for astrophysical objects, example: Search "Andromeda" (a famous galaxy) in SIMBAD
- for solar system bodies: https://vo.imcce.fr/webservices/ssodnet/?quaero (API)
- ongoing study for “observation facilities” (based on wikidata, joint action between IPDA and IVOA)

What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

Within the scope of the data centres and tool developers adhering to the standards of the various sub-domain international alliances, the main limitating / constraining factor is usually the **complexity of implementation** of the standards (mostly a human resource issue).

Sub-domain search interfaces are available and are using standard sub-domain semantic artefacts. However, **long tail data products usually not built using the sub-domain standards**.

Most tools are presenting semantic artefacts to users when data products have been discovered, but they are **not so much used to defined automated pipelines/processing**.

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

There is no overall catalogue for semantic artefacts
⇒ no catalogue mapping (some study were conducted, but never used operationally)
⇒ but there are **places listing vocabularies and models** (alliances website, FAIRsharing…)

There are resource registries:
searchable catalogues of data resources (or ancillary products) based on metadata standards (one for each alliance)

Example of study in *Space Physics*: http://wdcosf.fh-potsdam.de/espas
(vocabulary broker between Space Physics, Earth sciences, Astronomy, Climate… not maintained since 2019)

What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

Current semantic ecosystems are managed at a global scale (worldwide alliances), and are globally used by many actors (datacenters, developers…).

Semantic artefact catalogues not so much needed?
⇒ Except for science topics covering neighbouring fields.

For neighbouring fields: need for mapping of keywords and values.

Still many “niche”-sub-topics are not covered by existing semantic artefacts from international alliances.

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

Some crosswalks/mappings has been produced in the past years

- schema.org ⇔ SPASE [IHDEA] (RDA metadata crosswalk WG)
- EPNcore [IVOA] ⇔ SOLARNET [IHDEA] (EPN-RI-2024 project)
- EPNcore [IVOA] ⇔ SPASE [IHDEA] (EPN-RI-2024 project)
- EPNcore [IVOA] ⇔ PDS4 [IPDA] (EPN-RI-2020 project)

These mapping are between dictionary keywords, not on value terms.

Push for reusing common vocabularies for terms

Earth Science

(J-C Desconnets, C. Pierkot, V. Agazzi)

Which are the semantic artefacts that you develop or use in your scientific domain and how are they governed (developed and maintained)?

data model / knowledge model (ontology)

- geospatial community : OGC , ISO TC 211 <-> W3C
  - closed world : ISO, OGC Spécifications decline in
  - open world via W3C recommendations e.g SOSA/SSN,I-Adopt, DCAT, GEOSPARQL

onto-terminologies: large quantity of SKOS thesauri that deals with a subset of the domain

- overview of earth science : NASA Global Change Master Directory
- Nerc Vocabulary (Ocean), Sandre (water), EnvThes, Convention Climate forescat (climate), Taxref (species) and so on…
- terminology back bone for Data Terra research infracture
- https://docs.google.com/presentation/d/1cfRa2E0V_kr1h346nMt9CYI6FIENbuUhHvdhnKlPH2U/edit?usp=sharing (slides 24-> 34)

What limitations do you see in your scientific domain with respect to the use of semantic artefacts to describe and/or search/find research datasets?

- **content and access**
  - different level of fair maturity of thesaurii, ontology
  - high fragmentation of artifacts within platforms and a very high level of granularity of description
  - lack of semantic access and annotations services
- **use of semantic artefact in information system**
  - lack of skills in data manager spatial community to design and implement
  - limited implementations that support geographic semantic artefacts

Is there any type of semantic artefact catalogue in your scientific domain, where some (or all) of these semantic artefacts are catalogued?

- **statements**
  - weak number and partial one e.g NERC vocabulary for ocean domain
  - each data sharing platform designed and set up own thesaurii, ontology (as data terra center) without best practises
- **planned**
  - Implementation of an *EarthPortal* in the framework of FAIR-IMPACT based on Ontoportal technology

What limitations do you see in your scientific domain with respect to the use of semantic artefacts catalogues?

- **Limitations and needs**
  - difficulties to find existing artefacts and reuse it
  - lack of consistency of semantic artefacts
  - No mapping/crosswalks between semantic artefacts

Is there a strategy to deal with crosswalk and mappings between semantic artefacts in your scientific domain?

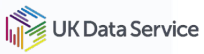
In data Terra research infrastructure

- state of the art of semantic artefact (thesaurii) in the differents earth communities (earth compartment)
- building consistency between semantic artefact with terminological backbone
- bests practices and recommendations to structure and annotate data with semantic artefacts

More generally

- Use of the *EarthPortal* to improve the mapping/crosswalk between semantic artefacts of the Earth Science communities

The Consortium



eosc | FAIR-IMPACT

Expanding FAIR solutions across EOSC

@fairimpact_eu /company/fair-impact-eu-project

Funded by
the European Union