

Deliverable 2.11

Review on the existing terminologies and terminology services



| | |
|-------------------------------|---|
| Deliverable no. | D2.11 |
| Work package | WP2 - ACCESS TO FACILITIES, FAIR DATA AND RELATED SERVICES |
| Intermediate Objective | IO2.1 |
| Deliverable type | <input checked="" type="checkbox"/> Document, report <input type="checkbox"/> Websites, patent filings, videos, etc. <input type="checkbox"/> Other: please specify |
| Dissemination level | <input checked="" type="checkbox"/> Public <input type="checkbox"/> Restricted |
| Estimated delivery (bimester) | B3 |
| First delivery date | 30/04/2023 |
| Final delivery date | 30/10/2023 |
| Authors (Partner-OU) | Di Muri Cristina, Muresan Nicoleta Alexandra, Raho Davide, Rosati Ilaria |
| Reviewed by | Carmela Cornacchia (CNR-IMAA) |
| Comments | |

INDEX

| | |
|---|--------------------|
| 1. LIST OF ACRONYMS | 4 |
| 2. GLOSSARY | 6 |
| 3. INTRODUCTION | 8 |
| 3.1 Scientific Context | 8 |
| 3.2 Deliverable Structure | 9 |
| 4. METHODS | 10 |
| 5. SEMANTIC ARTEFACTS AND REPOSITORIES: AN OVERVIEW | 12 |
| 6. SEMANTIC ARTEFACTS AND THEIR DOMAINS | 16 |
| 7. SEMANTIC ARTEFACTS ADOPTED BY THE NATIONAL RIs | 18 |
| 8. DISCUSSION AND FINAL REMARKS | 20 |

1. LIST OF ACRONYMS

ACTRIS: Aerosol, Clouds and Trace Gases Research Infrastructure

API: Application Programming Interface

ARDC: Australian Research Data Commons

BARTOC: Basic Register of Thesauri, Ontologies and Classifications

BODC: British Oceanographic Data Centre

CC: Creative Commons

CSV: Comma-Separated Values

DL: Description Logic

DO: Digital Object

eLTER: European Long-Term Ecosystem and socio-ecological Research infrastructure

EMBL-EBI: European Molecular Biology Laboratory - European Bioinformatics Institute

EMSO: European Multidisciplinary Seafloor and water-column Observatory

ERIC: European Research Infrastructure Consortium

FAIR: Findable, Accessible, Interoperable and Reusable

GFBio: German Federation for Biological data

GUI: Graphic User Interface

HTTP: Hypertext Transfer Protocol

IBISBA: Industrial Biotechnology Innovation and Synthetic Biology Acceleration

ICOS: Integrated Carbon Observation System

ITINERIS: Italian Integrated Environmental Research Infrastructures System

JERICO: Joint European Research Infrastructure of Coastal Observatories

JSON: JavaScript Object Notation

JSON-LD: JavaScript Object Notation for Linked Data

LOTERRE: Linked Open TERminology REsources

LOV: Linked Open Vocabularies

MIT: Massachusetts Institute of Technology

MMI-ORR: Marine Metadata Interoperability Ontology Registry and Repository

MOD: Metadata for Ontology Description and publication

M2M: Machine to Machine

NERC: Natural Environment Research Council (of UK)

NVS: NERC Vocabulary Server

OBO: Open Biological and Biomedical Ontologies

OLS: Ontology Lookup Service

OU: Operative Unit

OWL: Web Ontology Language

RDF: Resource Description Framework

RDFS: Resource Description Framework Schema

REST: Representational State Transfer

RI: Research Infrastructure

RPC: Remote Procedure Call

SIOS: Svalbard Integrated Arctic Earth Observing System

SKOS: Simple Knowledge Organisation System

SOAP: Simple Object Access Protocol

SPARQL: Simple Protocol and RDF Query Language

TTL: Turtle Graphics Data Format

URL: Uniform Resource Locator

WP: Work Package

XLS: Microsoft Excel spreadsheet

XML: Extensible Markup Language

2. GLOSSARY

Authority file: Authority files are lists of terms that are used to control the variant names for an entity or the domain value for a particular field. Examples include names for countries, individuals, and organisations. Non preferred terms may be linked to the preferred versions. This type of KOS generally does not include a deep organisation or complex structure. The presentation may be alphabetical or organised by a shallow classification scheme. A limited hierarchy may be applied to allow for simple navigation, particularly when the authority file is being accessed manually or is extremely large.

Categorisation scheme: A categorisation scheme is a set of controlled terms whose entities are divided into “buckets” or broad topic levels. Some examples provide a hierarchical arrangement of numeric or alphabetic notation to represent broad topics. These types of KOSs lack the explicit relationships presented in a thesaurus.

Controlled vocabulary: A controlled vocabulary is a normalised, restricted list of terms for a specific use or context. Thesauri and taxonomies are types of controlled vocabularies, but not all controlled vocabularies are thesauri or taxonomies.

Gazetteer: A gazetteer is a list of place names. Traditional gazetteers have been published as books or have appeared as indexes to atlases. Each entry may also be identified by feature type, such as river, city, or school. Geospatially referenced gazetteers provide coordinates for locating the place on the earth’s surface. The term gazetteer has several other meanings, including an announcement publication such as a patent or legal gazetteer. These gazetteers are often organised using classification schemes or subject categories.

Glossary: A glossary is an alphabetical list of terms in a particular domain of knowledge with the definitions for those terms.

Ontology: An ontology is a formal version of a thesaurus where relations are described using a formal system such as Description Logic (DL) to mathematically classify individuals of classes and properties.

Semantic artefact: 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.

Semantic Resources...

Semantic registry: A semantic registry is a catalogue that contains metadata about semantic artefacts.

Semantic repository: A semantic repository is defined in this recommendation as a service that stores and offers access to both the metadata of semantic artefacts and their content, *i.e.* offers search and access to get individual terms (including their metadata) both for humans and for machines.

Subject heading: A subject heading is a set of controlled terms to represent the subjects of items in a collection. Subject headings can be extensive and cover a broad range of subjects; however, their structure is generally very shallow, with a limited hierarchical

structure. In use, subject headings tend to be coordinated, with rules for how they can be joined to provide concepts that are more specific.

Taxonomy: A taxonomy is a controlled vocabulary with a hierarchical structure used to classify things or concepts. Terms within a taxonomy have relations to other terms (parent/broader term, child/narrower term).

Term/class/concept: A term/class/concept is an individual element with a unique semantic interpretation, represented with a unique identifier.

Thesaurus: A thesaurus is essentially a controlled vocabulary following a standard structure, where all terms have relationships of three kinds to each other: hierarchical (broader term/narrower term), associative (related term), and equivalent (use/used for or see/ seen from). Some terms in thesauri might have additional explanatory notes, such as scope notes (brief explanations about the coverage of the term or of how it should be used in indexing) or history notes. Thesauri are defined in the [ISO25964](#).

3. INTRODUCTION

The deliverable 2.11 is produced within the framework of the ITINERIS project and it is part of the activity 2.4 of the Work Package (WP) 2 concerning the access to facilities, FAIR data and services provided by the 22 Italian Research Infrastructures (RIs) involved in the project. This deliverable belongs to the intermediate objective of bimester 3 and it is produced under the responsibility of the Operative Unit (OU) of the National Research Council, Research Institute on Terrestrial Ecosystems (CNR-IRET) in Lecce.

The overarching objective of this deliverable is to provide an extensive review of existing terminologies and terminology services relevant for the four environmental domains considered within ITINERIS (*i.e.* marine, atmosphere, geosphere landsurface and terrestrial biosphere). In addition, this report provides suggestions about terminologies and terminology services that could be adopted within ITINERIS and discusses the existing gaps and the envisaged implementations of the project. This review provides an initial assessment of the resources to be harvested by ITINERIS terminology service. Furthermore, this deliverable highlights the types of resources that are still lacking and that should be implemented to support the interoperability of the Digital Objects (DOs) provided by the 22 RIs and to enhance the interdisciplinarity across the four different environmental domains of the project.

From now on, within this deliverable and, more in general, within the framework of the WP 2 activities, “terminology” will be replaced by “semantic artefact” in agreement with the international recommendations outlined by the [FAIRsFAIR](#) and [FAIR-IMPACT](#) projects (Le Franc *et al.*, 2022), by the [EOSC Task Force](#) on semantic interoperability (EC, 2021), and by the Metadata for Ontology Description and publication ([MOD](#)).

3.1 Scientific Context

The deliverable 2.11 constitutes a key milestone to enhance the data semantic interoperability of Italian environmental RIs. Semantic interoperability allows the transmission of the meaning of data and it is critical for supporting collaborative data-intensive research and enabling the generation of novel scientific knowledge (Karam *et al.*, 2018). Semantic interoperability represents one of the fundamental pillars of FAIR (Findable, Accessible, Interoperable, Reusable; Wilkinson *et al.*, 2016) data and of all DOs. Specifically, it encapsulates the second principle of Interoperability (I2) stating that (meta)data should be described by controlled vocabularies documented and resolvable using globally unique and persistent identifiers ([FAIR Principles](#)).

A key component of semantic interoperability is the adoption of semantic artefacts, providing a framework for conflict resolution between terms developed independently across different domains. Currently, different terms may have been used to describe the same concept or, alternatively, the same term may be used to express several concepts within different domains and/or disciplines, thereby impeding interoperability. Semantic artefacts are machine-actionable formalisations of concepts that enable the discovery, the integration and the reuse of information by both humans and machines (Le Franc *et al.*, 2022). They may have a broad range of formalisations, from loose sets of terms such as glossaries and categorisation schemes to higher-order logic constructs such as thesauri and ontologies (Corcho *et al.*, 2023; check Zeng, 2008 for an extensive overview) and are therefore built using different standard models (*e.g.* RDFs, OWL, SKOS) and serialisation formats (*e.g.* XML, XML Schema, JSON, RDF/XML, OWL/XML, JSON-LD, Turtle,

N-Triples, *etc.*) (Le Franc *et al.*, 2022). Often, these semantic artefacts are stored and shared by means of two types of catalogues: *i.* semantic registries defined as metadata catalogues of semantic artefacts; and *ii.* semantic repositories that store and offer access to both the metadata of semantic artefacts and their content, each providing a mixture of functionality (Le Franc *et al.*, 2022; Corcho *et al.*, 2023). By accessing and interrogating them, users can discover the most appropriate standard terms to be used for the semantic annotation of their (meta)data. Semantic repositories may also include the semantic mappings between concepts and indicate the relations among them, with the goal of helping the users to understand not only the meaning of the terms, but also the context in which they can be used. In addition, semantic repositories are used to overcome the language barriers by favouring multilingualism (Benis *et al.*, 2022). Semantic registries and repositories are essential to facilitate the discoverability and the access to existing semantic artefacts, to support their management and use and, in some cases, they also provide editing tools. The formalisation and digitalisation of concepts provided by these services allows the technological enhancement of a number of digital disciplines including (meta)data management, information retrieval, recommendation systems and, more in general, the development of artificial intelligence systems and of semantic web technologies (Middleton *et al.*, 2004).

3.2 Deliverable Structure

This report is divided into eight sections, including the acronyms and the glossary sections, and of which this introduction is the third one. Section 4 describes the methodological approach adopted for semantic artefacts collection and analysis. Section 5 provides a general overview and description of the compiled semantic artefacts and of the semantic catalogues relevant for ITINERIS. Section 6 provides a more detailed description, for each environmental domain, of the main characteristics of the semantic artefacts found during this review. Section 7 focuses on the resources currently adopted by the RIs involved in the project and, in section 8, a gap analysis and a needs analysis is carried out to highlight what it is required within the project in order to fulfil the interdisciplinarity and the interoperability of (meta)data provided by the national RIs.

References are reported at the end of the document (section 9) and Annex 1, which includes information about the semantic artefacts considered, is provided as an attachment.

4. METHODS

The systematic review carried out to compile an extensive list of semantic artefacts relevant for ITINERIS was mainly conducted through the interrogation of 17 semantic repositories and registries listed in Table 1. These catalogues were selected as they offer access to a wide array of semantic artefacts inherent to different environmental domains and, in general, for environmental research. The catalogues with less than 500 semantic artefacts (Tab. 1) were thoroughly interrogated with the aim of selecting the most suitable resources for ITINERIS. On the other hand, the catalogues with a larger number of semantic artefacts (> 500; Tab. 1), were explored by either using APIs or SPARQL interfaces ([sparql](#)), or by performing manual queries and filtering by specific keywords or relevant subjects and/or domains. In addition to this preliminary investigation, the Google search engine (<https://www.google.it/>) was interrogated to retrieve further semantic artefacts not included in the repositories and registries listed in Table 1. The Google search was performed using keywords such as “semantic” or “semantic artefacts” together with domain specific keywords such as “marine” or “atmosphere” or “biosphere” or “geosphere” or “landsurface”.

At the end of this search process, only the semantic artefacts relevant for ITINERIS were selected and listed in a spreadsheet (see Annex 1) with the following associated information whereby attributes’ names and definitions are aligned, whenever possible, to the [MOD v. 2.0](#) schema:

- **Title**: A name given to the resource;
- **Acronym**: Often used as an identifier within some ontology platforms such as BioPortal or OBO Foundry;
- **Type**: The nature or genre of the resource (*e.g.* ontology, thesaurus, *etc.*);
- **Theme**: A main category of the resource. A resource can have multiple themes (*i.e.* marine, atmosphere, biosphere, geosphere);
- **Date Issued**: Date of formal issuance of the resource;
- **Version**: The version number of the resource;
- **Access URL**: A URL of the resource that gives access to a distribution of the dataset. *E.g.* landing page, feed, SPARQL endpoint;
- **Format**: The file format of the resource (*e.g.* XML, XML Schema, JSON, RDF/XML, OWL/XML, JSON-LD, Turtle, N-Triples, *etc.*);
- **Included in data catalog**: A data catalog which contains this dataset;
- **Licence**: A legal document giving official permission to do something with the resource;
- **Source accessed on**: The resource is related to a source which was originally accessed or consulted on the given date as part of creating or authoring the resource.

Beyond this review of semantic artefacts, this report took advantage of a survey, launched in March 2023 in the framework of WP2 Activity 2.3, to gather the semantic artefacts adopted and/or produced and maintained by the ITINERIS RIs. The general aim of the survey was to collect the FAIR-enabling best practices currently adopted by the national RIs involved in the project. From the survey, the semantic resources reported by the RIs were gathered and analysed. Further analyses of the results of this survey are still in progress, and they will also feed the deliverable 2.7 “State of the art review of FAIR-enabling best practices”.

The data analysis and data visualisation, shown in the following sections, was performed in the R-Studio interface, R engine version 4.3.0 (R Core Team, 2021). Data manipulation was carried out with the packages *dplyr* v. 1.1.2 (Wickam *et al.*, 2021), *tidyverse* v. 2.0.0 (Wickam *et al.*, 2019), and *reshape* v. 2.1 (Wickam, 2007). Topic modelling analysis was carried out using the full title of all semantic artefacts and the packages *tm* v. 0.7-11 (Feinerer and Hornik 2023) and *wordcloud* v. 2.6 (Fellows, 2018). To clean the text, the functions *removePunctuation*, *removeNumbers*, *stripWhitespace*, and *content_transformer* within *tm* were used, respectively, to eliminate punctuation marks, numbers, extra white spaces and to convert all letters to lowercase. The *tm* function *removeWords* was applied to remove stopwords and common words such as semantic, ontology, thesaurus, vocabulary, glossary, terms, terminology as well as project specific words such as NERC, Argo, SeaDataNet *etc.* Lastly, the Venn diagram and the sankey diagram were obtained respectively using the packages *networkD3* v. 0.4 (Allaire *et al.*, 2017) and *ggvenn* v. 0.1.10 (Yan, 2023).

Table 1. List of semantic repositories and registries, in alphabetical order, interrogated to compile this review. The table includes their full names and acronyms, the type of catalogue (repository or registry), the total number of semantic artefacts included and their online Uniform Resource Locators (URLs).

| Repository name (Acronym) | Type | No. of semantic artefacts | Online URL |
|---|------------|---------------------------|---|
| AberOWL | Repository | 1,422 | http://aber-owl.net/#/ |
| AgroPortal | Repository | 150 | https://agroportal.lirmm.fr/ |
| Basic Register of Thesauri, Ontologies & Classifications (BARTOC) | Registry | 3,429 | https://bartoc.org/ |
| BioPortal | Repository | 1,044 | https://bioportal.bioontology.org/ |
| Bioregistry | Registry | 1,650 | https://bioregistry.io/ |
| NERC Vocabulary Server (NVS) | Repository | 292 | https://vocab.nerc.ac.uk/ |
| EcoPortal | Repository | 25 | https://ecoportal.lifewatch.eu/ |
| FAIRsharing | Registry | 833 | https://fairsharing.org/ |
| German Federation for Biological data (GFBio) | Repository | 29 | https://terminologies.gfbio.org/ |
| I-ADOPT Catalogue of Terminologies | Registry | 84 | https://i-adopt.github.io/terminologies |
| Linked Open TERminology REsources (LOTERRRE) | Repository | 65 | https://www.loterre.fr/presentation/ |
| Linked Open Vocabularies (LOV) | Repository | 812 | https://lov.linkeddata.es/dataset/lov/ |
| Marine Metadata Interoperability Ontology Registry and Repository (MMI-ORR) | Repository | 330 | https://mmisw.org/ |
| Ontobee | Repository | 262 | https://ontobee.org/ |
| Open Biological and Biomedical Ontology (OBO) Foundry | Registry | 185 | https://obofoundry.org/ |
| EMBL-EBI Ontology Lookup Service (OLS) | Repository | 242 | https://www.ebi.ac.uk/ols/index |
| Research Vocabularies Australia (ARDC) | Repository | 443 | https://vocabs.ardc.edu.au/ |

5. SEMANTIC ARTEFACTS AND REPOSITORIES: AN OVERVIEW

Overall, a total number of 540 semantic artefacts were found including ontologies ($N = 199$), thesauri ($N = 153$), glossaries ($N = 145$), gazetteers ($N = 18$), categorisation schemes ($N = 13$), subject headings ($N = 7$), and authority files ($N = 5$). Based on their complexity level, these semantic artefacts were expressed in different formats (*e.g.* OWL/XML, RDF/XML, TTL, JSON, JSON-LD) and most of them were available in multiple formats (Annex 1). Among these resources, 225 provide semantic artefacts exclusively for the terrestrial biosphere domain, 60 for the geosphere landsurface domain, 48 for the marine domain, and four for the atmosphere domain (Fig. 1). In addition, 143 semantic artefacts cover all domains and 60 concern multiple domains (at least two).

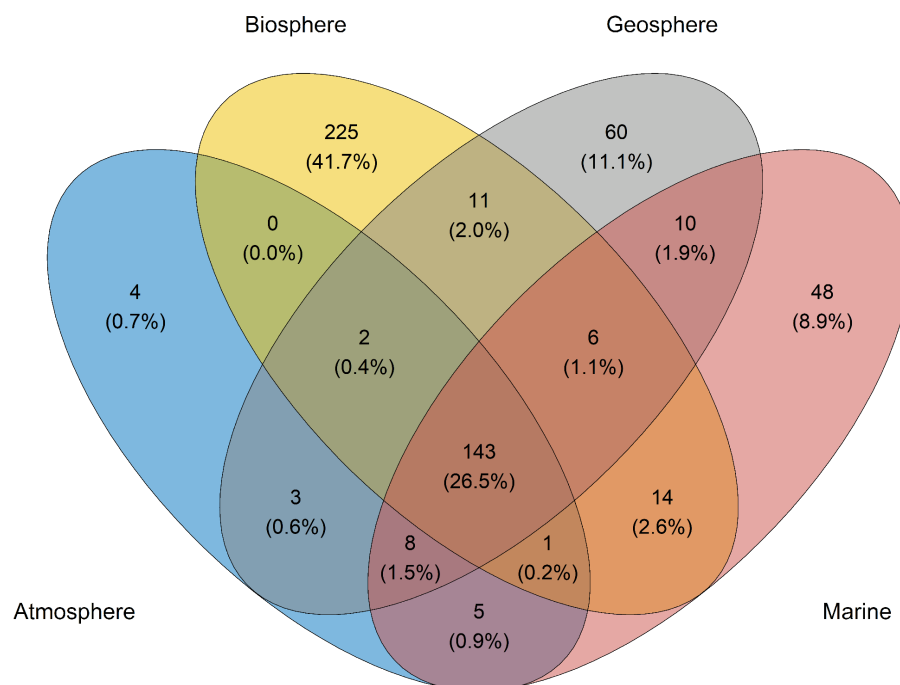


Figure 1. Venn diagram displaying the number of shared and unique semantic artefacts among the different ITINERIS environmental domains.

As for the specific topics covered, figure 2 shows a word cloud image of the most common topics covered by the semantic artefacts as resulted from the topic modelling analysis. The majority of the semantic artefacts describe parameters ($freq = 43$), anatomies ($freq = 24$), environments ($freq = 23$), developmental stages ($freq = 19$), phenotypes ($freq = 15$), biological entities ($freq = 14$), coastal ($freq = 14$) and marine entities ($freq = 14$), methods ($freq = 13$), traits ($freq = 12$), observations ($freq = 11$), plants ($freq = 11$), taxonomies ($freq = 11$), data ($freq = 10$), sensors ($freq = 10$), and units ($freq = 10$).

- Editing/managing existing resources through the use of internal or external integrated web platforms. For example, EcoPortal integrates the collaborative, multilingual, open source platform Vocbench 3 for the editing and development of SKOS thesauri, OWL ontologies and RDF datasets in general (Stellato *et al.*, 2020).

All such functionalities allow the DOs discovery, accession, curation and annotation through different interfaces: Graphic User Interface (GUI, WEB-GUI) and Machine-to-Machine interface (M2M). Through the GUI, users can perform some or all the actions listed above whereas the M2M, more often known as Application Programming Interface (API) ensures the interoperability of the resources through automatised and defined schemes. The most common APIs for the evaluated repositories were RESTful (Representational State Transfer) APIs and the SOAP (Simple Object Access Protocol) APIs. The RESTful API has an architecture based on standard communication protocols (*i.e.* HTTP) and the REST requests are also referred to as stateless, as each request made by the client to the server contains all is needed to fulfil the request. REST resources are reached via access points represented by URIs. The types of requests that can be enabled are GET (to obtain information), POST (to create a new resource), PUT (to update a resource) and DELETE (to delete the resource). The server replies are in the form of HTTP and also contain other information on the status of the service. The SOAP APIs are based on an XML (eXtensible Markup Language) messaging protocol called SOAP. They follow the Remote Procedure Call (RPC) paradigm, which is based on the invocation of remote procedures on the server. Communication takes place through SOAP messages encapsulated in XML format and communicated through a HTTP protocol, which contains the information about the command to be executed, the type of request and the output. Some repositories also provide a SPARQL endpoint. This is a query language used to extract information from RDF-based datasets and data can be queried using a wide range of operators. A SPARQL Endpoint is a server that has an interface to execute SPARQL queries. Users can send SPARQL queries to an endpoint to obtain the desired results. Query results can be returned in RDF format or in other formats such as JSON or XML.

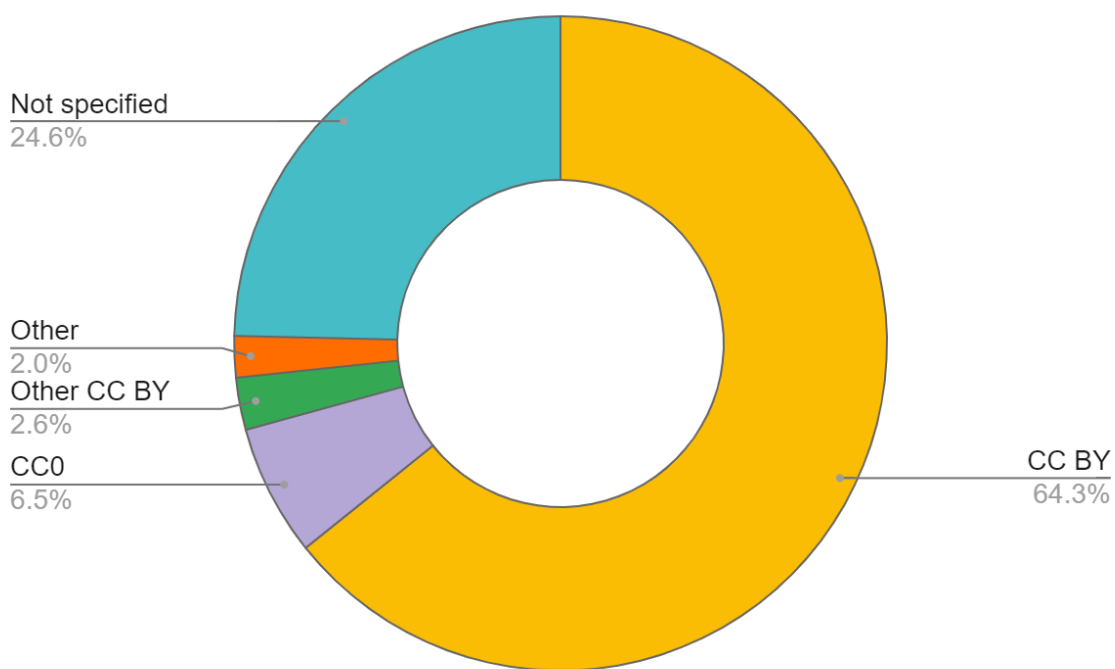


Figure 3. Type and percentage of semantic artefacts licences.

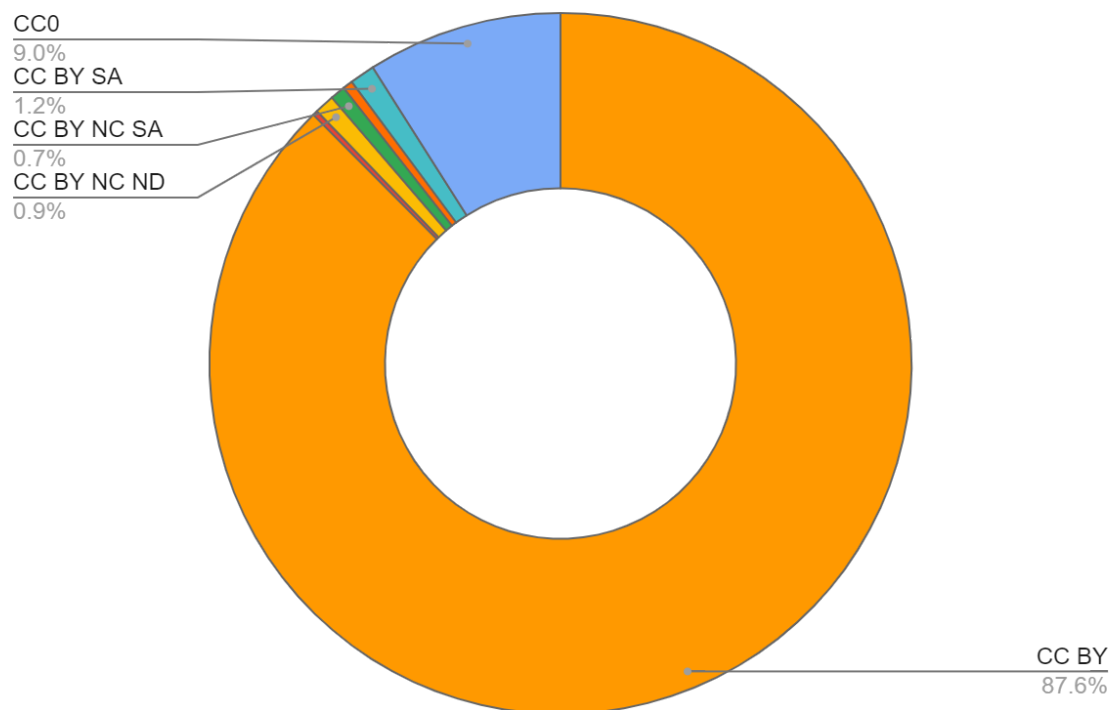


Figure 4. Type and percentage of the semantic artefacts published under Creative Commons.

6. SEMANTIC ARTEFACTS AND THEIR DOMAINS

This section enhances the analysis of the semantic artefacts by domain. As highlighted above, of the 540 semantic artefacts identified, 143 were classified as relevant to all four environmental domains (Fig. 1; Tab. 2). The majority of resources in this category are thesauri ($N = 64$; Tab. 2) included in the NERC vocabulary server (Annex 1). These vocabularies cover a wide range of topics and have broad applicability, encompassing biological, chemical, and physical entities, as well as instruments, measurements, and Standard Operating Procedures (SOPs; Annex 1). A further set of 60 semantic artefacts were identified as relevant to two or more domains (Fig. 1; Tab. 2). Most of these resources shared by multiple domains ($N = 44$; Fig. 1; Tab. 2) focused on the marine domain in conjunction with another domain. This is likely due to the abundance of semantic artefacts sourced from the NERC vocabulary server, which is managed by the British Oceanographic Data Centre (Annex 1).

As for the semantic artefacts assigned to a single domain, the terrestrial biosphere domain includes the highest number of resources ($N = 225$; Fig. 1, 3; Tab. 2) consisting of 156 ontologies, 37 glossaries, 19 thesauri, 10 categorisation schemes and 3 subject headings (Tab. 2) that cover a wide range of topics from biological observations to organisms' traits. Additionally, many resources target specific taxonomic groups or species, with some representing common model organisms such as *Caenorhabditis elegans*, *Xenopus spp.*, *Danio rerio*, and laboratory mice (Annex 1).

As for the remaining semantic artefacts assigned to a single domain, 60 of them were assigned to the geosphere landsurface domain, predominantly in the form of thesauri ($N = 37$; Tab. 2; Fig. 5). The marine domain included 48 semantic artefacts published mostly as glossaries ($N = 33$; Tab. 2), whereas the atmosphere domain included four resources including two glossaries, one ontology and one thesaurus (Fig. 1, 3; Tab. 2).

Table 2. Number and type of semantic artefacts per domain. The table also includes the number of resources shared by all or multiple domains.

| Domain | Authority file | Categorisation scheme | Gazetteer | Glossary | Ontology | Subject heading | Thesaurus |
|--------------------------------|----------------|-----------------------|-----------|----------|----------|-----------------|-----------|
| All | 3 | 1 | 2 | 45 | 25 | 3 | 64 |
| Atmosphere | 0 | 0 | 0 | 2 | 1 | 0 | 1 |
| Atmosphere-Biosphere-Geosphere | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Atmosphere-Biosphere-Marine | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Atmosphere-Geosphere | 0 | 0 | 0 | 0 | 2 | 0 | 1 |
| Atmosphere-Geosphere-Marine | 0 | 0 | 0 | 4 | 1 | 0 | 3 |
| Atmosphere-Marine | 0 | 0 | 0 | 3 | 0 | 0 | 2 |
| Biosphere | 0 | 10 | 0 | 37 | 156 | 3 | 19 |
| Biosphere-Geosphere | 0 | 0 | 0 | 2 | 2 | 0 | 7 |
| Biosphere-Geosphere-Marine | 0 | 1 | 0 | 2 | 1 | 0 | 2 |
| Biosphere-Marine | 0 | 0 | 1 | 4 | 3 | 0 | 6 |
| Geosphere | 0 | 0 | 9 | 8 | 6 | 0 | 37 |
| Geosphere-Marine | 0 | 0 | 3 | 5 | 0 | 0 | 2 |
| Marine | 2 | 1 | 3 | 33 | 2 | 0 | 7 |

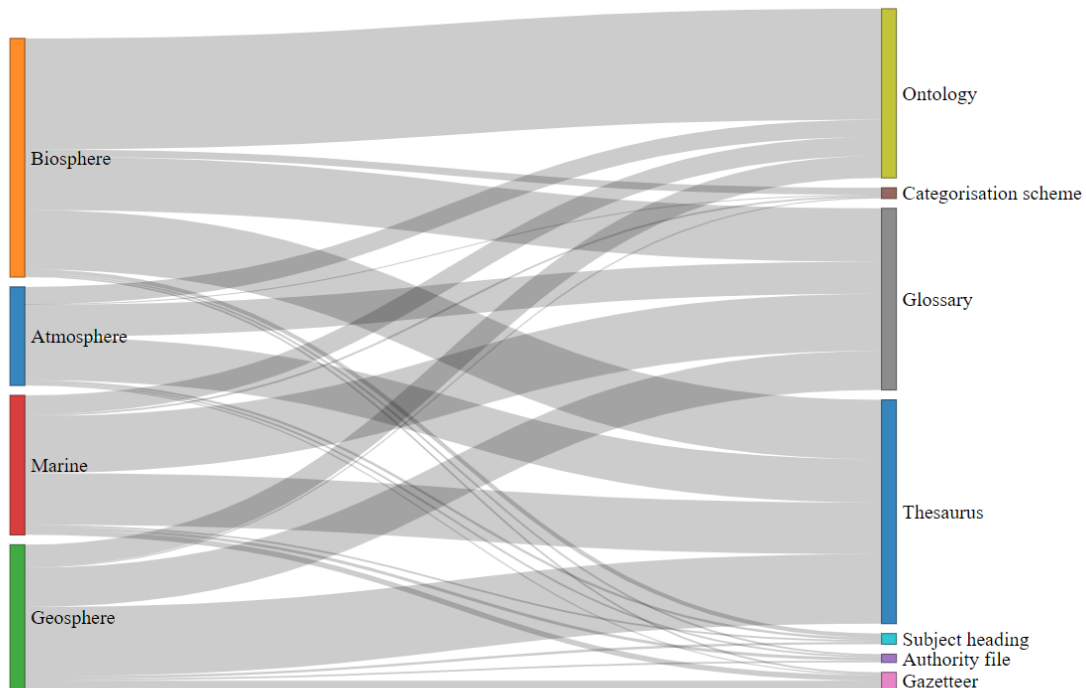


Figure 5. Distribution of types of semantic artefacts available per domain.

7. SEMANTIC ARTEFACTS ADOPTED BY THE NATIONAL RIs

From the survey results it is inferred that only 11 out of the 22 ITINERIS RIs are using semantic artefacts to ensure the interoperability of their (meta)data (Tab. 3). Four out of the 11 RIs, *i.e.*, ACTRIS, ICOS, LifeWatch ERIC and eLTER RI, are also semantic artefacts providers and use both external and internal semantic resources that are developed and maintained by them (Tab. 3).

The ACTRIS Vocabulary includes 1,777 concepts and uses the I-ADOPT model (Magagna *et al.*, 2021) for atomising the variable names into part of themselves whereby the latter ones consist in a controlled list of terms.




The internal ICOS ontology supports the metadata schema of the ICOS Carbon Portal.

The LifeWatch ERIC semantic artefacts, available through EcoPortal, are SKOS thesauri, maintained by the Italian node of the RI, that focus on organisms' traits with the addition of an upper ontology used within the infrastructure.

The eLTER RI vocabularies include two SKOS thesauri, namely EnvThes and eLTER_CS. The latter one is used internally to manage the eLTER community, whereas EnvThes has a broader application and it is used to describe concepts of long-term ecological research, monitoring and experiments.

Among the semantic artefacts provided by external servers, the BODC Vocabularies and, in particular, the ones describing parameters and device categories are the most commonly used resources (Tab. 3). It should be noted that DiSSCo (not listed in Tab. 3) stated that the semantic artefacts that are going to be used within the infrastructure are currently under construction and they are going to be available on the GitHub page of the infrastructure (<https://github.com/DiSSCo>), whereas IBISBA indicated the repository of semantic artefacts used to retrieve concepts used within the infrastructure (*i.e.* EMBL-EBI OLS), but not the specific resources used (Tab. 3).

Table 3. List of semantic artefacts (full name and hyperlink) adopted by the Research Infrastructures.

| Research Infrastructure | Adopted semantic artefacts | Domain |
|---|--|------------|
|  | ACTRIS Vocabulary Climate and Forecast Standard Names | Atmosphere |
|  | BODC Parameter Usage Vocabulary BODC-approved data storage units SeaDataNet device categories SeaVoX Platform Categories SeaVoX Device Catalogue | Marine |
|  | Argo parameter codes | Marine |

| | | |
|--|--|--|
|  <p>European Research Infrastructure Consortium</p> | <p>Ontology of Integrated Carbon Observation System (ICOS) Creative Commons Rights Expression Language DCMI Metadata Terms Semantic Sensor Network Ontology The PROV Ontology Sensor, Observation, Sample, and Actuator (SOSA) Ontology VANN: A vocabulary for annotating vocabulary descriptions An RDF/OWL vocabulary for representing spatial information Schema.org Vocabularies</p> | <p>Atmosphere Marine Biosphere</p> |
|  | <p>Fish Traits Thesaurus Zooplankton Traits Thesaurus Phytoplankton Traits Thesaurus LifeWatch ERIC Upper Ontology I-ADOPT Framework Ontology Darwin Core BODC Parameter Usage Vocabulary Environmental Thesaurus</p> | <p>Biosphere</p> |
|  | <p>SeaDataNet Parameter Discovery Vocabulary</p> | <p>Marine</p> |
|  | <p>eLTER Vocabularies BODC Vocabularies</p> | <p>Marine Biosphere</p> |
|  | <p>Climate and Forecast Metadata Convention</p> | <p>Atmosphere Marine</p> |
| <p>Laura Bassi</p>  | <p>SeaVoX Device Catalogue SeaDataNet Device Categories SeaDataNet Platform Categories</p> | <p>Marine</p> |
|  | <p>EMBL-EBI Ontology Lookup Service</p> | <p>Biosphere</p> |
|  | <p>BODC Parameter Usage Vocabulary BODC-approved data storage units MEDATLAS Parameter Usage Vocabulary</p> | <p>Marine</p> |

*This table includes only the list of semantic artefacts used by the RIs involved in ITINERIS. Data, metadata and semantic standards provided in the survey were excluded as they are not considered relevant for this review.

8. DISCUSSION AND FINAL REMARKS

This deliverable provides an initial extensive review of the semantic artefacts that are going to be harvested by the terminology service developed within ITINERIS to support the interoperability of the (meta)data provided by the 22 RIs and to enhance the interdisciplinarity between the different environmental domains considered within the project.

In this review, the majority of compiled semantic artefacts belong to the terrestrial biosphere domain ($N = 255$; Fig. 1; Tab. 2). Although many resources within this domain describe organisms' observations and traits, the high number of semantic artefacts found might have been also determined by the domains' classification established within ITINERIS. For example, the resources within the field of hydrology have been considered part of the terrestrial biosphere domain according to the project's internal categorisation scheme. Compared to the other environmental domains, atmosphere is by far the one with the lowest number of semantic artefacts both if considered on its own ($N = 4$; Fig. 1; Tab. 2) and if accounting for the semantic resources shared with other domains ($N = 19$; Fig. 1; Tab. 2). It should be noted, however, that many semantic artefacts herein classified as belonging to all environmental domains focused on climate. Such resources included cross-cutting terms and many of them described atmospheric-related concepts. For the purpose of this deliverable, indeed, the semantic artefacts have been assigned to a single domain only and only if all terms therein included described concepts pertaining to that specific domain. The Authors' personal decision on the resources categorisation undoubtedly played a role in influencing the quantity of semantic artefacts identified within the different environmental domains.

Despite the high number of semantic artefacts compiled in this review, the majority of them focus on a relatively few topics, generally applicable to all domains, such as environmental parameters, models, instruments and measurement units (Fig. 2). In addition, and in line with the higher number of resources belonging to the terrestrial biosphere domain (Fig. 1; Tab. 2), the remaining common topics are associated to anatomies and biological entities, with many resources focusing on developmental stages, plants and phenotypes (Fig. 2). This result might indicate a possible lack of semantic artefacts describing other topics of interest for the project ITINERIS and for the RIs involved. This possible gap should be further investigated and adequately addressed in order to suit the specific needs of the ITINERIS RIs.

Most of the semantic artefacts with an available licence were published under the CC BY licence (Fig. 3, 4). All such licences allow the sharing (copy and redistribution) and the adaptation/transformation of the resources for any purpose. Hence, such semantic artefacts could be then fetched by the terminology service developed within ITINERIS. For a few resources ($N = 133$; Fig. 3), the licence was not explicitly stated by the creators and, in case they will be considered needful for the ITINERIS RIs, the contact person in charge of these semantic artefacts should be reached.

The results of the survey indicated that only 11 out of the 22 ITINERIS RIs use semantic artefacts and four of them are also providers of semantic artefacts. Differences in the use of practices to ensure the semantic interoperability between RIs could be associated with the level of maturity of the different RIs. In addition to this, RIs are not required to adopt strategies and standards for making their (meta)data interoperable. In the context of

ITINERIS, and in alignment with other European projects in which some ITINERIS RIs are involved, e.g. [FAIR-IMPACT](#), [ENVRI-FAIR](#), one of the key objectives is to define and provide guidelines, at national level, for the adoption of practices to ensure an accurate representation and annotation of the (meta)data and, ultimately, guarantee their interoperability. This deliverable represents the initial stride in establishing these guidelines and their associated recommendations. One of these recommendations would be the suggestion of [EcoPortal](#) (Kechagioglou *et al.*, 2021), managed by LifeWatch ERIC, as a candidate repository of semantic artefacts for those resources that are going to be eventually developed within ITINERIS. In fact, this review has revealed that LifeWatch ERIC is the only RI, among those involved in ITINERIS, that manages a repository of semantic artefacts providing the possibility to create, edit, and publish FAIR semantic artefacts. Moreover, EcoPortal collects semantic artefacts within the biodiversity and ecosystems domain, hence, accounting for the wide range of environmental domains covered, it could be considered appropriate to store and publish the new semantic artefacts generated within the project.

This review was essential to understand the key features that the ITINERIS terminology service should have to support the (meta)data interoperability within ITINERIS. One of the main features of the terminology service is to provide a unique access point to all the relevant environmental semantic artefacts queried by the service. This will entail the integration of semantic artefacts with different types of formalisation and available in a range of different formats with varying degrees of semantic interoperability. A further essential requirement of the ITINERIS terminology service will be the inclusion of an automatic AI-guided mapping service to determine correspondences and interlinks between terms of different resources in order to avoid semantic inconsistencies and redundancy. Those alignments are fundamental for the efficient integration and access to the DOs provided within ITINERIS. The ITINERIS terminology service should additionally enable an efficient semantic annotation and provide to the users the most suitable terms to describe their (meta)data. Through the use of logical reasoning capabilities facilitated by ontologies, annotated DOs can be accessed with greater efficiency in specific application scenarios. By extracting additional information from ontological statements, the search capabilities can be expanded significantly beyond a limited set of keywords and, such extension, empowers users with a more valuable and precise data discovery process (Karam *et al.*, 2016).

While carrying out this review, it has been noted that the semantic artefacts and their catalogues exhibit various degrees of FAIRness. It is widely acknowledged that numerous semantic artefacts have a limited level of FAIRness due to challenges related to their discoverability and accessibility (e.g. not included in any repository) or interoperability (e.g. lack of mappings) (Goldfarb and LeFranc, 2017). Although the primary objective of this deliverable did not include a FAIRness assessment of the compiled semantic artefacts, it is essential to consider this aspect when contemplating their reuse. Corcho *et al.* (2023) performed an extensive investigation of the maturity level of 26 repositories of semantic artefacts and developed a model to facilitate their future assessments. The study's findings indicated varying maturity levels across the repositories considered. The FAIRness assessment of all the distinct resources included within these repositories would also be essential as, in fact, based on the Authors' personal observations, certain resources included in this review may not meet the FAIR criteria. For example, when the Authors have queried the AberOWL repository through its API to extract the number of semantic

artefacts available, the repository's server responded with a total number of 1,422 resources (Tab. 1). However, only 1,051 of them were accessible, whereas the remaining were "Unloadable" ($N = 212$) or "Incoherent" ($N = 6$) or an "Unknown Error" ($N = 153$) was generated during the query. To evaluate the FAIRness of semantic artefacts, [AgroPortal](#) and [EcoPortal](#) have included a FAIR assessment tool with the capacity to assign a FAIR score to each resource as well as to the entire repository (Amdouni *et al.*, 2022). This or other tools could be also used for the FAIRness assessment of the semantic artefacts collated in this review before integrating them within the ITINERIS terminology service.

It should be noted that whilst the reasonable collective effort carried out in this exercise (Annex 1), the compiled list of semantic artefacts should not be considered exhaustive in either depth or breadth. To ensure the semantic interoperability of (meta)data provided by the ITINERIS RIs throughout the lifetime of the project and beyond, this list should be constantly updated as new semantic artefacts of interest are identified and adopted. Furthermore, the involvement of data managers of the ITINERIS RIs in different domains will be key to complete this initial review and keep up to date the list of semantic artefacts collected.

9. REFERENCES

Allaire JJ, Ellis P, Gandrud C, Kuo K, Lewis BW, Owen J *et al.*, 2017. D3 JavaScript Network Graphs from R. R package version 0.4. URL: <https://cran.r-project.org/package=networkD3>

Amdouni E, Bouazzouni S, Jonquet C, 2022. O'FAIRE: Ontology FAIRness Evaluator in the AgroPortal semantic resource repository. ESWC 2022 - 19th Extended Semantic Web Conference. pp. 89-94, 10.1007/978-3-031-11609-4_17

Benis A, Grosjean J, Billey K, Montanha G, Dornauer V, Crişan-Vida M *et al.*, 2022. Medical informatics and digital health multilingual ontology (MIMO): a tool to improve international collaborations. *International Journal of Medical Informatics*, 167:104860. DOI: 10.1016/j.ijmedinf.2022.104860

Corcho O, Ekaputra FJ, Heibi I, Jonquet C, Micsik A, Peroni S, Storti E, 2023. A maturity model for catalogues of semantic artefacts. *arXiv preprint arXiv:2305.06746*

European Commission, Directorate-General for Research and Innovation, Corcho O, Eriksson M, Kurowski K, *et al.*, 2021. *EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Available at: <https://data.europa.eu/doi/10.2777/620649>

Feinerer I and Hornik K, 2023. tm: Text Mining Package. R package version 0.7-11. URL: <https://CRAN.R-project.org/package=TM>

Fellows I, 2018. wordcloud: Word Clouds. R package version 2.6. URL: <https://CRAN.R-project.org/package=wordcloud>

Goldfarb D, Le Franc Y, 2017. Enhancing the Discoverability and Interoperability of Multi-Disciplinary Semantic Repositories. In S4BioDiv@ ISWC.

- Karam N, Müller-Birn C, Gleisberg M, Fichtmüller D, Tolksdorf R, Güntsch A, 2016. A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. *Datenbank-Spektrum*, 16, 195-205. DOI: 10.1007/s13222-016-0231-8
- Le Franc Y, Bonino L, Koivula H, Parland von-Hessen J, Pergl R, 2022. D2.8 FAIR Semantics Recommendations Third Iteration. Version 1. DOI: 10.5281/zenodo.6276576
- Magagna B, Rosati I, Stoica M, Schindler S, Moncoiffe G, Devaraju A *et al.*, 2021. The I-ADOPT Interoperability Framework for FAIRer data descriptions of biodiversity. *arXiv preprint arXiv:2107.06547*
- Middleton SE, De Roure D, Shadbolt NR, 2004. Ontology-based Recommender Systems. In: Staab S, Studer R (eds) *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-24750-0_24
- Pan American Health Organization, 2021. Introduction to Semantic Interoperability. Digital transformation toolkit, Knowledge Tools. Available at: https://iris.paho.org/bitstream/handle/10665.2/55417/PAHOEIHIS21023_eng.pdf?sequence=1&isAllowed=y#:~:text=Semantic%20interoperability%20is%20the%20ability.of%20the%20meaning%20of%20data.
- R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>
- Stellato A, Fiorelli M, Turbati A, Lorenzetti T, Van Gemert W, Dechandon D *et al.*, 2020. VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5), 855-881. DOI: 10.3233/SW-200370
- Wickham H, 2007. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), URL: <http://www.jstatsoft.org/v21/i12/paper>
- Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L *et al.*, 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
- Wickham H, François R, Henry L, Müller K, 2021. dplyr: A Grammar of Data Manipulation. R package version 1.0.7. <https://CRAN.R-project.org/package=dplyr>
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Mons B *et al.*, 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9. DOI: 10.1038/sdata.2016.18
- Yan L, 2023. Draw Venn Diagram by 'ggplot2'. R package version 0.1.10. URL: <https://cran.r-project.org/web/packages/ggvenn/index.html>
- Zeng M, 2008. Knowledge Organization Systems (KOS). *Knowledge Organization*, 35, 160-182. DOI: 10.5771/0943-7444-2008-2-3-160.