



## Deliverable 2.13

# Description of The Terminology Interoperability Framework Architecture (1st release)



Deliverable no.	D2.13
Work package	WP2 - ACCESS TO FACILITIES, FAIR DATA AND RELATED SERVICES
Intermediate Objective	IO2.2
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)	B5
First delivery date	31/08/2023
Final delivery date	29/02/2024
Authors (Partner-OU)	Raho Davide, Rosati Ilaria
Reviewed by	Di Muri Cristina Muresan Alexandra Nicoleta
Comments	Due to the nature of the deliverable, this document shows two release dates: a first release as overseen by project and a final release updated to the latest information available. This document may undergo several updates and releases as the evolving technologies.

## INDEX

<a href="#">1. LIST OF ACRONYMS</a>	<a href="#">5</a>
<a href="#">2. GLOSSARY</a>	<a href="#">6</a>
<a href="#">3. Introduction</a>	<a href="#">8</a>
<a href="#">3.1. Interoperability</a>	<a href="#">8</a>
<a href="#">3.1.1. Semantic Interoperability</a>	<a href="#">9</a>
<a href="#">4. Terminology Interoperability Framework purpose</a>	<a href="#">10</a>
<a href="#">4.1. Scope, readership and usage of TIF</a>	<a href="#">11</a>
<a href="#">5. Background and Context</a>	<a href="#">12</a>
<a href="#">6. Principles of TIF</a>	<a href="#">13</a>
<a href="#">6.1. Integrability</a>	<a href="#">13</a>
<a href="#">6.2. Scalability</a>	<a href="#">13</a>
<a href="#">6.3. Internationalisation</a>	<a href="#">13</a>
<a href="#">6.4. Modularity</a>	<a href="#">14</a>
<a href="#">6.5. Automation</a>	<a href="#">14</a>
<a href="#">6.6. Reusability</a>	<a href="#">14</a>
<a href="#">6.7. User centricity</a>	<a href="#">15</a>
<a href="#">7. Core Components of the TIF</a>	<a href="#">15</a>
<a href="#">7.1. Terminology Models</a>	<a href="#">15</a>
<a href="#">7.2. Mapping and Alignment Techniques</a>	<a href="#">16</a>
<a href="#">7.3. Relevant Standards</a>	<a href="#">16</a>
<a href="#">7.3.1. Languages and formats</a>	<a href="#">16</a>
<a href="#">7.3.2. (Meta)data Standards</a>	<a href="#">18</a>
<a href="#">7.4. Relevant Data access Technologies</a>	<a href="#">19</a>
<a href="#">8. Architecture Design and Implementation Strategies</a>	<a href="#">19</a>
<a href="#">8.1. Application Architecture</a>	<a href="#">20</a>
<a href="#">8.1.1. Presentation Layer</a>	<a href="#">20</a>
<a href="#">8.1.2. Application Layer</a>	<a href="#">20</a>
<a href="#">8.1.3. Data storage and management</a>	<a href="#">21</a>
<a href="#">8.2. Meta-Data Component</a>	<a href="#">21</a>
<a href="#">8.3. Data Integration and Exchange Components</a>	<a href="#">22</a>
<a href="#">8.4. Security and Privacy Considerations</a>	<a href="#">22</a>
<a href="#">9. Use Case - ITINERIS Terminology Service</a>	<a href="#">22</a>
<a href="#">9.1. Terminology Service Requirements</a>	<a href="#">24</a>
<a href="#">9.2. Terminology Service Functionalities</a>	<a href="#">25</a>
<a href="#">9.2.1. Collection of Semantic Artefact Sources</a>	<a href="#">25</a>
<a href="#">9.2.2. Harvesting</a>	<a href="#">26</a>
<a href="#">9.2.2.1. Collection of Semantic Artefacts</a>	<a href="#">26</a>
<a href="#">9.2.2.2. Terminological Collection</a>	<a href="#">26</a>
<a href="#">9.2.3. Terminological Alignment (Lexicographic Mapping)</a>	<a href="#">26</a>

<a href="#">9.2.4. Ontology of Mapping and Linked Data</a>	<a href="#">27</a>
<a href="#">9.2.5. Networking (Linked Data)</a>	<a href="#">27</a>
<a href="#">9.2.6. Diffing</a>	<a href="#">28</a>
<a href="#">9.2.7. Semantic Annotation</a>	<a href="#">28</a>
<a href="#">9.2.8. Storage of Mappings and Linked Data</a>	<a href="#">29</a>
<a href="#">9.2.9. Search for Semantic Artefacts</a>	<a href="#">29</a>
<a href="#">9.2.10. Access to Semantic Artefacts</a>	<a href="#">30</a>
<a href="#">9.2.11. Terminological Search</a>	<a href="#">30</a>
<a href="#">9.2.12. Ranking of Terms in the Terminological Search Results Set</a>	<a href="#">30</a>
<a href="#">9.2.13. Search for Semantically Annotated Informational Objects</a>	<a href="#">31</a>
<a href="#">9.2.14. Dereferencing, Publishing, Editing</a>	<a href="#">31</a>
<a href="#">9.2.15. Creation and Management of Users and Roles</a>	<a href="#">32</a>
<a href="#">10. Conclusion</a>	<a href="#">32</a>
<a href="#">11. References</a>	<a href="#">33</a>

## 1. LIST OF ACRONYMS

**API:** Application Programming Interface

**CRUD:** Create, Read, Update, Delete

**CSS:** Cascading Style Sheets

**DCAT:** Data CATalogue Vocabulary

**DO:** Digital Object

**DwC:** Darwin Core

**EML:** Ecological Metadata Language

**FAIR:** Findable, Accessible, Interoperable, Reusable

**GUI:** Graphic User Interface

**HTML:** HyperText Markup Language

**HTTP:** HyperText Transfer Protocol

**IO:** Intermediate Object

**ITINERIS:** ITalian INTe grated Environmental Research Infrastructures System

**JS:** JavaScript

**JSON:** JavaScript Object Notation

**JSON-LD:** JavaScript Object Notation for Linked Data

**MOD:** Metadata for Ontology Description and publication

**M2M:** Machine to Machine

**OU:** Operative Unit

**OWL:** Web Ontology Language

**PROV-O:** Provenance Ontology

**RDF:** Resource Description Framework

**RDFa:** Resource Description Framework in attributes

**RDFS:** Resource Description Framework Schema

**REST:** Representational State Transfer

**RI:** Research Infrastructure

**SI:** Semantic interoperability

**SKOS:** Simple Knowledge Organisation System

**SOAP:** Simple Object Access Protocol

**SPARQL:** Simple Protocol and RDF Query Language

**SPOC:** Single Point Of Contact

**SSSOM:** Simple Standard for Sharing Ontology Mappings

**TIF:** Terminology Interoperability Framework

**TS:** Terminology Service

**TTL:** Terse RDF Triple Language

**URI:** Uniform Resource Identifier

**URL:** Uniform Resource Locator

**WP:** Work Package

**XML:** Extensible Markup Language

## 2. GLOSSARY

**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.

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

**ITINERIS:** Project for building the Italian Hub of Research Infrastructures in the environmental scientific domain for the observation and study of environmental processes in the atmosphere, marine domain, terrestrial biosphere, and geosphere, providing access to data and services and supporting the Country to address current and expected environmental challenges.

**List:** Limited sets of terms in some sequential order.

**Metadata:** is structured information which describes an information resource (data).

**Ontology:** An ontology is a formal description of knowledge 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 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.

**Terminology Service:** Overarching expression to refer to any kind web service and application serving terminology content.

**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 2.13 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 environmental Research Infrastructures (RIs) involved in the project. This deliverable, together with del. 2.1 - User strategy; del. 2.4 - ITINERIS HUB; del. 2.7 - State of the Art review of FAIR-enabling best practises; del. 2.8 Fair Implementation Profiles (FIPs); del. 2.12 - FAIR terminologies, belongs to the Intermediate Objective (IO) 2.2 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 describe the architecture of the Terminology Interoperability Framework (TIF). Del. 2.13 will focus on a set of principles, requirements, strategies and suggestions to drive its development as well as the development of the Terminology Service (TS) and associated functionalities. Moreover, during the development and integration stages of the software components within and beyond the ITINERIS TS, further integrations are foreseen to make the framework document more detailed and re-usable.

#### 3.1. INTEROPERABILITY

Interoperability is "the capability of two or more networks, systems, devices, applications, or components to externally exchange and readily use information securely and effectively or for a device to operate as an equivalent substitute for some other device"<sup>1</sup>. Interoperability refers to the ability of computer systems or softwares to access, exchange, integrate and cooperatively use data in a coordinated manner. Interoperability can operate at local or at distributed level, within and across organisational, political and jurisdictional boundaries, to provide timely and seamless portability of information<sup>2</sup>. Interoperability is crucial whenever independent and different systems, managed by various jurisdictions cooperate together in an efficient and effective way<sup>3</sup>. The adoption of standards, i.e. architectures, frameworks, terminologies, datatypes and coding is required to reach the interoperability<sup>4</sup>.

Interoperability is a pillar property in a world that day by day is projecting towards a total digitalisation, where different systems communicate and collaborate effectively.

Although there are several types of interoperability<sup>5</sup>, generally speaking this can be organised into five popular different layers:

---

<sup>1</sup> IEEE Std 2030™

<sup>2</sup> <https://www.himss.org/resources/interoperability-healthcare>

<sup>3</sup> Gottschalk, 2009

<sup>4</sup> Jepsen et al., 2010

<sup>5</sup> Ford et al., 2007

- Foundational (technical) that establishes the requirements for securely exchange data within and across systems, it is typically associated with hardware and software components and communication protocols<sup>6</sup>;
- Syntactic that establishes formats, syntax and organisation of data exchange;
- Semantics that focuses on the ability to interpret the meaning of the exchanged data. Semantic interoperability uses common underlying models and codification of data through the use of data elements with standard definitions, providing shared understanding and meaning. The use of common ontologies, metadata schemas, and controlled vocabularies ensures that the meaning of data is maintained and understood evenly across systems;
- Organisational that includes governance, policy, social, legal and organisational considerations to facilitate communication and use of data within and between entities. It represents the capacity of organisations to work together towards common goals. It involves governance and organisational consideration, sharing processes, policies and procedures and often requires legal agreements or collaborations enabling shared consent, trust and integrated end-user processes in seamless workflows.

This document focuses mainly on Semantic Interoperability referring to the other layers whenever needed.

### 3.1.1. Semantic Interoperability

In information systems, semantic interoperability (SI) is the ability of different systems to interpret and utilise data coherently at a meaningful level<sup>7</sup>. The SI ensures uniform comprehension of information, overcoming challenges posed by linguistic and conceptual differences. At its core, SI focuses on the meaning and context of data, distinguishing itself from syntactic interoperability, which is concerned with the structure of data, and technical interoperability, which deals with the communication protocols between systems. Terminology interoperability involves the use of standardised vocabularies, terms, and definitions to facilitate coherent communication and data exchange. This concept is especially critical in fields that rely heavily on data sharing and collaboration, such as healthcare<sup>8</sup> and environmental research. It enables the integration of data from various sources, fostering a comprehensive understanding of environmental issues. This concept is

---

<sup>6</sup> Rezaei et al., 2014

<sup>7</sup> Pan American Health Organization (2021). Introduction to Semantic Interoperability. (<https://iris.paho.org/handle/10665.2/55417>).

<sup>8</sup> Chatterjee et al., 2022

not just a technical requirement but a vital enabler for collaborative research and the generation of new scientific insights<sup>9</sup>. SI represents the third fundamental principle of FAIR (Findable, Accessible, Interoperable, Reusable) data and Digital Objects (DOs; [FAIR Principles](#)<sup>10</sup>). The Interoperability principle advocates for (meta)data to be described by controlled vocabularies that are well-documented and associated to globally unique and persistent identifiers. In the real world, different terms are used to describe the same concept or, sometimes, the same term could be used to express several concepts within different domains and/or disciplines, thereby hindering the interoperability. Terminology interoperability refers to the capacity of different information systems, databases, and digital platforms to consistently use and interpret standard sets of terms and vocabularies. It ensures that terms have the same meaning and implications, regardless of the system or the stakeholder engaging with the data. In the context of environmental research, this means the sharing of data and knowledge among scientists, policymakers, and widely speaking all stakeholders. In other words, data related to climate change, biodiversity, pollution, conservation, and sustainability can be shared, interpreted, and reused coherently across different geographical locations, scientific disciplines, and technological frameworks. Semantic artefacts are machine-actionable and machine-readable formalisations of concepts that enable the discovery, the integration and the reuse of information by both humans and machines<sup>11</sup>. The role of semantic artefacts like ontologies and controlled vocabularies is pivotal in the Terminology Interoperability, serving as the backbone for a shared understanding of data. Ontologies and controlled vocabularies provide a common framework for representing and sharing meanings. Standards such as RDF (Resource Description Framework), OWL (Web Ontology Language), and SKOS (Simple Knowledge Organization System) are instrumental in structuring and linking data semantically, thus facilitating a coherent and unified interpretation of information across different systems. These technologies enable the creation of extensive knowledge frameworks that span various domains, ensuring that data is not only accessible but also interpretable on a broad scale.

#### 4. TERMINOLOGY INTEROPERABILITY FRAMEWORK PURPOSE

The main purposes of the ITINERIS Terminology Interoperability Framework (TIF) are to inspire Italian RIs and researchers, provide guidance to RIs and contribute to fostering the research process. The TIF implementation aims to harmonise the access to and the information exchange about terminologies in the ITINERIS environment to seamlessly integrate disparate terminologies. By standardising terminology usage and facilitating interoperability, TIF not only streamlines the communication but it also enhances data consistency and accuracy, ultimately optimising research processes and fostering synergy

---

<sup>9</sup> Karam et al., 2016

<sup>10</sup> <https://www.go-fair.org/fair-principles/>

<sup>11</sup> Le Franc et al., 2022

across diverse domains within the ITINERIS ecosystem. In synthesis, the TIF is a commonly shared approach to deliver ITINERIS services in an interoperable manner.

In the context of ITINERIS, the TIF embodies a structured ensemble of principles and conventions delineated to facilitate the interoperability across independent systems and to guide the development of new software applications specifically geared towards terminology interoperability. TIF will delineate a blueprint for organised structure, elucidating how various components shall interconnect and interface with one another to ensure seamless interoperability. Within the ITINERIS ecosystem, developers will benefit from a repository of reusable components and modules tailored to address the intricacies of terminology management and integration. Pre-configured components such as widget components and API web services are built to alleviate the web developing processes. Moreover, TIF offers mechanisms for extending and customising these components to harmonise with the idiosyncrasies of diverse application contexts. Furthermore, TIF advocates for adherence to standardised conventions and protocols, leaving a high grade of liberty in the development of federated applications.

In essence, the TIF serves as an indispensable scaffold for the development of services semantically related aimed at facilitating the interchange and harmonisation of terminologies.

Thus, TIF establishes standards for the correct functioning of the ITINERIS TS leading to an easy and seamless flow in the discovery, access, exchange and harvesting of terminologies.

#### 4.1. Scope, readership and usage of TIF

The scope of TIF is to harmonise information exchanged across services making them efficiently and effectively collaborative. The ITINERIS TS will be developed following TIF recommendations and requirements.

In particular, the TIF can be considered as a guide for different stakeholders including: *i.* RIs seeking to standardise data exchange formats for an enhanced public service delivery; *ii.* for developers working on intuitive and interoperable software solutions; *iii.* data curators aiming to improve the accessibility and linkage of archival resources; and *iv.* researchers sharing reproducible data across various domains and infrastructures. The TIF includes guidance to the integration and harmonisation of terminological resources, and facilitates the use of semantic technologies for improved data discoverability, interoperability, and reuse, ensuring that knowledge is accurately represented and easily accessible in digital ecosystems.

The TIF enables the harmonised exchange of information about terminologies through the Terminology service system, harmonising the information exchange interfaces of pre-existing RIs systems and third-party services. The TIF is defined and will be tailored

to enable the logic and the workflow for retrieving terminologies from different sources within and beyond the ITINERIS project, such as external repositories and independent terminology services. In addition, the TIF will make terminologies available within the ITINERIS ecosystem providing endpoints ready to be used by machine or by humans through a user interface.

The ITINERIS Interoperability Framework is designed to address the RIs' needs which could be heterogeneous in relation to the different domains of research. For this reason, to enable an efficient and effective consumption of terminologies, TIF will consider the needs of making available tailored slices over terminology. Moreover, the TIF establishes rules and requisites for mediating requests for terminology updates between users and providers.

## 5. BACKGROUND AND CONTEXT

For the design and development of a Service Oriented Architecture it is necessary to have a thorough knowledge of the domain of interest and the needs and expectations of stakeholders. Moreover, in a framework that connects several different information systems, a deep understanding of those systems is needed in order to reach good levels of interoperability across the different interoperability layers, from the technical one up to the organisational one.

Within the ITINERIS WP2, several key deliverables have been produced. These deliverables are crucial for the development of the ITINERIS TIF.

Deliverable D2.7 - "State of the Art review of FAIR-enabling best practices" (Activity 2.3), provides a comprehensive analysis of FAIR implementation practices by presenting a general overview of the strategies and methodologies employed by the ITINERIS RIs to adhere to FAIR principles. The document seeks to facilitate the exchange of best practices among RIs, thereby enhancing the implementation of FAIR-enabling solutions within the national RIs landscape in the environmental sector.

Deliverable D2.8 - "Fair Implementation Profiles (FIPs) - First release" (Activity 2.3), encompasses a review of existing FIPs, accessible via the FIP Wizard, relevant to the ITINERIS project. The FIP essentially serves as a formal declaration of selected technology choices, namely FAIR Enabling Resources (FERs), by a FAIR Implementation Community (FIC) to address one or more FAIR Guiding Principles. FIPs are designed to monitor the progress and convergence of FAIR data services within RIs by evaluating the commonality of FERs between them.

Lastly, Deliverable D2.11 - "Review on the existing terminologies and terminology services" (Activity 2.4), offers an exhaustive review of terminologies and terminology services pertinent to ITINERIS purposes and needs. It assessed the terminologies and services currently utilised by RIs, identified existing gaps, and discussed potential

implementations for ITINERIS. This review also provided an initial evaluation of the resources that the forthcoming terminology service within ITINERIS will need to harvest, highlighting the types of resources required to support the interoperability of DOs provided by the 22 ITINERIS RIs and to foster interdisciplinarity across the project's environmental domains.

## 6. PRINCIPLES OF TIF

### 6.1. Integrability

Following the needs of RIs and their unicity, a set of tools or interfaces need to be conceptualised for the exposition of terminology services. Services have to be usable in different ways: through a set of web services that expose the system's functionalities, and through front-end components that encapsulate these services, offering visual interaction elements. This encapsulation strategy will help developers and administrators of different web portals to directly integrate services in their informatic systems. Moreover, the RIs and, more in general, the host sites, will be able to develop custom front-end elements that invoke the system's web services, or integrate the already available ones. Integrability, for specific purposes, will ensure also bidirectional data transmission between the TSs and existing data and metadata management systems (e.g., service catalogues). By prioritising interfaces and protocols that promote easy integration, it will ensure that diverse terminological systems can communicate effectively, fostering a unified semantic environment. This approach not only enhances data exchange but also supports the collaborative evolution of semantic networks.

### 6.2. Scalability

The system's back-end design shall consider an increasing demand for services over time, ideally with a vertical scaling model. It is important that systems have the ability to efficiently adapt and manage an increasing number of terminologies and integration needs. This ensures seamless performance and reliability, even as data volumes, user numbers, and transaction rates grow. By designing for scalability, the architecture supports evolving semantic web technologies and facilitates broad, cross-domain interoperability, making it a cornerstone for sustainable, long-term semantic integration strategies. In a multi-modular sourcing strategy which gets semantic artefacts from several external sources, an increasing effort will be required in cross-monitoring individual sources and maintaining correspondence maps.

### 6.3. Internationalisation

TIF architecture will be designed to be adaptable to different languages and regions without the needs of engineering changes. Following international standards, systems will

be designed to be accessible by international organisations and people. At the user level, web systems developed will include configurable systems for changing the user interface language, and for easily adding new translation languages (e.g., via a translation file for the labels of graphic components). At the system level, Internationalisation should be guaranteed through predefined mappings of semantic resources.

#### 6.4. Modularity

Modularity ensures flexibility and scalability, allowing for the seamless integration of new components or the updating of existing ones without disrupting the overall system. This principle facilitates tailored solutions to specific domain needs, enabling a dynamic and adaptive approach to manage terminological resources and services. By fostering an ecosystem where modules can be developed, replaced, or enhanced independently, modularity supports a sustainable, evolving ecosystem that can accommodate technological advancements and changing user requirements.

#### 6.5. Automation

The correct functioning of Interoperability will be ensured with a high degree of workflow automation and use of sophisticated software algorithms and processes to streamline the management, mapping, and integration of semantic resources. This includes automated harvesting of semantic artefacts, dynamic updating of terminological mappings, and seamless integration of new data sources. This aspect needs to be declinated in every system functionalities, describing the degree of automation and the desired mode of operation. By leveraging automation, TIF aims to reduce manual efforts, enhance efficiency, and ensure the timely and accurate alignment of terminological resources across different systems and domains. This foundational principle supports the scalability and adaptability of semantic interoperability solutions, facilitating a more cohesive and interconnected digital ecosystem.

#### 6.6. Reusability

This principle advocates for the creation of adaptable, general-purpose semantic assets that can be effectively integrated into multiple projects, enhancing the efficiency and scalability of semantic interoperability efforts. By prioritising reusability, the framework ensures that terminological resources contribute to a sustainable ecosystem, facilitating broader adoption and fostering innovation in data management and sharing practices.

Moreover, within the ITINERIS project, reusability will also be encouraged through the early design, the development and the subsequent release of codes, tools and functionalities easily adaptable to different circumstances and needs.

## 6.7. User centricity

User is a crucial point in the development of the TIF and its derived TS. User centricity approach ensures that the framework not only facilitates semantic interoperability but also promotes user experience by making terminological resources more accessible, intuitive, and useful for individuals across different domains. The framework aims to support more effective data discovery, annotation, and reuse, thereby promoting broader adoption and more meaningful engagement with semantic technologies.

Users shall be identified in as many different roles and with as many different permissions taking in consideration their different needs, such as discoverability for guest or researcher and managing needs for Administrators. A User-focused approach can ensure an easy and precise use of services, avoiding reluctance and limitations in the use of terminology systems.

User centricity will be guaranteed also through different solutions which transversally consider all other principles. For example, user login will be supported by a single sign-on system to ensure a seamless experience.

## 7. CORE COMPONENTS OF THE TIF

The core components of a TIF will include the Terminology Management, Standardisation and Governance and Interoperability mechanisms. Terminology Management involves the systematic collection, organisation and maintenance of terms and their meanings; Standardisation and Governance considers guidelines and procedures across several different platforms and systems; Interoperability mechanisms boasts tools and protocols that enable different systems to understand and use the terminologies effectively. This includes APIs, web services, and other middleware solutions.

### 7.1. Terminology Models

Terminology models are conceptual models that represent the structure, relationships, and attributes of terms within a domain and serve to the structure and interoperability of semantic web technologies. They provide a formal representation of knowledge domains through structured semantic artefacts, facilitating the sharing and integration of data across different systems and platforms. Examples of terminology models include controlled vocabularies, which are standardised lists of terms with specific meanings within a given context, or thesauri which include synonyms and hierarchical relationships among terms. Ontologies are graph-based models that define any type of relationships (properties) between terms (classes). These models enable precise communication between systems, enhance data discoverability, and support the development of intelligent applications capable of understanding and processing complex datasets.

## 7.2. Mapping and Alignment Techniques

The main scope of TIF is to harmonise access to terminologies between various different systems with their own organisation system and metadata and metadata schema. To do so, Mappings and Crosswalks are critical for linking classes of Ontologies, concepts of thesauri and meanings between metadata fields. These operations are time consuming and often need a deep understanding of the domain of interest. There are several approaches to do so: manual mapping is the most accurate one, performed by domain experts that identify and link equivalent or related terms across systems, is time-consuming and not scalable; a semi-automated approach, where software tools find likely associations that need to be verified, to be corrected or to be accepted by domain expert; lastly, automated Mapping Tools utilise algorithms to identify potential matches based on terms similarities, definitions, and context.

## 7.3. Relevant Standards

Terminology models can be expressed in different standard languages (e.g. RDF, SKOS, and OWL) which facilitate the description and linking of data in a format understandable to both humans and machines. These standards allow the accurate mapping and interpretation of data, overcoming language and disciplinary barriers and promoting greater interoperability and understanding between different information systems.

### 7.3.1. Languages and formats

The Resource Description Framework<sup>12</sup> (RDF) is the standard model for information interchange in the semantic web. RDF was designed to provide a common way to describe information that can be read and understood by computer applications but it is not designed to be displayed on the web and it is not particularly human readable either. The structure of RDF is predicated upon triples, that consider a resource, a property and a property value also known as subject, predicate, and object. These triples facilitate the linking of web resources and the specification of relationships among them. The RDF documents are written in XML (RDF/XML, with extension \*.rdf).

Other syntaxes and language formats used to improve the informativeness and the readability are N-triples<sup>13</sup>, Turtle<sup>14</sup> and Notation3<sup>15</sup>. The N-triples (with extension \*.nt) represents the concrete syntax of the RDF triples. N-Triples syntax is extremely simple and rigorous. Each line of the document represents a single triple RDF, with subject, predicate, and object separated by spaces and terminated by a point. There are no abbreviations or constructs to reduce verbosity, which makes it less readable for humans but easily processed by software.

---

<sup>12</sup> <https://www.w3.org/RDF/>

<sup>13</sup> <https://www.w3.org/TR/n-triples/>

<sup>14</sup> <https://www.w3.org/TR/turtle/>

<sup>15</sup> <https://www.w3.org/TeamSubmission/n3/>

The Turtle format (with extension \*.ttl) represents a more human readable version of RDF, employing a compact and simplified syntax to represent data in a clear manner. It permits the use of prefixes to reduce the verbosity of the URIs and the aggregation of triples that have the same subject. Turtle language can be embedded in HTML documents.

Notation3 (with extension \*.n3), designed to be more compact and readable than XML/RDF, extends RDF with additional features for expressing logic and rules.

The Web Ontology Language<sup>16</sup> (OWL) is a representation language for ontologies within the Semantic Web. Ontologies serve as essential tools for the conceptual modelling of information and the relationships among entities. The file extensions for OWL documents include ".owl" and ".rdf", given that ontologies are frequently represented using RDF.

RDF Schema<sup>17</sup> (RDFS) is a semantic extension of RDF that provides mechanisms for describing groups of related resources and the relationships between these resources. It is used to create vocabularies (or schemas) that describe the properties and classes of RDF. RDFS provides a basic framework for semantic descriptions of resources, including the definition of classes and properties, enabling more sophisticated and structured data modelling compared to RDF. Instead RDF Schema provides the framework to describe application-specific classes and properties.

Simple Knowledge Organisation System<sup>18</sup> (SKOS) is designed to represent different models including thesauri, classification schemes, subject heading systems, and taxonomies within the framework of the Semantic Web. SKOS is implemented in RDF and provides a common data model that allows diverse systems to be used and shared across different applications. It provides a standardised way to express the structure and content of concept schemes, including concepts, their labels, semantic relationships (like broader, narrower, related concepts), and documentation (such as definitions and notes).

Resource Description Framework in attributes<sup>19</sup> (RDFa) represents a technology that facilitates the embedding of RDF metadata directly into HTML web pages. This enables search engines and intelligent agents to more effectively comprehend the significance of the information contained within a web page. It does not possess a specific file extension, as it is directly incorporated into the HTML code of web pages.

JSON for Linking Data<sup>20</sup> (JSON-LD) is a format designed to facilitate the embedding of RDF data within JSON documents. This simplifies the representation of semantically structured data in web applications. JSON-LD objects can be included in files bearing the ".jsonld" extension.

---

<sup>16</sup> <https://www.w3.org/OWL/>

<sup>17</sup> <https://www.w3.org/2001/sw/wiki/RDFS>

<sup>18</sup> <https://www.w3.org/2004/02/skos/>

<sup>19</sup> <https://www.w3.org/2001/sw/wiki/RDFa>

<sup>20</sup> <https://www.w3.org/2001/sw/wiki/JSON-LD>

### 7.3.2. (Meta)data Standards

Darwin Core (DwC) is a standard of Biodiversity Information Standards (TDWG) organisation, maintained by the Darwin Core Maintenance Interest Group. It comprises a glossary of terms (which could be referred to as properties, elements, fields, columns, attributes, or concepts in different scenarios) designed to support the dissemination of data on biological diversity. It achieves this goal by offering identifiers, labels, and definitions<sup>21</sup>.

Ecological Metadata Language (EML) is a standardised vocabulary with an XML markup for describing research data, adopted in earth and environmental sciences and increasingly across other disciplines. It supports data documentation, preservation, and sharing, evolving through community efforts. EML enables comprehensive data package identification, method description, and precise semantic annotations, catering to the data documentation needs of researchers for open sharing and preservation<sup>22</sup>.

The Data Catalog Vocabulary<sup>23</sup> (DCAT) is designed to facilitate interoperability between data catalogues published on the Web. It can be used to describe dataset and data services. It enables the representation of data in a standard model, making it easier to discover, understand, and consume information about datasets, their distributions, and related resources.

DCAT Application Profile for data portals in Europe<sup>24</sup> (DCAT-AP) is a Linked Open Data specification for metadata made by the European Commission. Based on DCAT, it supports standardised dataset descriptions, increasing interoperability between all data portals in Europe.

Provenance Ontology<sup>25</sup> (PROV-O) based on OWL2 language focuses on representing provenance information, documenting the origin and transformations of data. It emphasises traceability and understanding causal relationships between entities involved in a process.

Metadata for Ontology Description<sup>26</sup> (MOD) is designed as an OWL ontology to describe metadata information for ontologies.

The Simple Standard for Sharing Ontology Mappings<sup>27</sup> (SSSOM) metadata standard for describing semantic mappings. It offers a comprehensive metadata term catalogue for describing mappings and their provenance, enhancing interoperability and curation with existing OWL and SKOS predicates.

---

<sup>21</sup> <https://dwc.tdwg.org/>

<sup>22</sup> Jones et al., 2019.

<sup>23</sup> <https://www.w3.org/TR/vocab-dcat-3/>

<sup>24</sup> <https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe>

<sup>25</sup> <https://www.w3.org/TR/prov-o/>

<sup>26</sup> <https://w3id.org/mod>

<sup>27</sup> <https://mapping-commons.github.io/sssom/spec/>

#### 7.4. Relevant Data access Technologies

The SPARQL Protocol and RDF Query Language<sup>28</sup> is an interrogation language specifically designed for RDF data. It is used to extract information from RDF graphs and perform complex searches on semantic data. SPARQL language adopts a Turtle style syntax and its queries are often stored in files with the ".rq" or ".sparql" extension.

Representational State Transfer (REST) Web APIs embody a set of architectural principles for designing web services. REST architecture permits CRUD-like operations on resources by GET, POST, PATCH, PUT, DELETE, which are uniquely identified through URIs/URLs. REST is predicated on a stateless client-server communication style, wherein each request contains all the necessary information to be understood by the server. Responses can be formatted in JSON, XML, or other media formats, making REST flexible and widely adopted for the development of modern and scalable APIs.

### 8. ARCHITECTURE DESIGN AND IMPLEMENTATION STRATEGIES

The architecture of a TIF will encompass several layers, each crucial for facilitating effective communication, data exchange, and understanding across varied systems and applications.

The architecture of a TIF is multi-faceted, requiring careful consideration of user interaction, core functionalities, data management, integration capabilities, and security measures. This can be achieved following and adhering to standards and best practices. Security and privacy are paramount, necessitating robust mechanisms to protect sensitive information and comply with regulatory requirements. ITINERIS TIF will follow recommendations that can be applied from the European Interoperability Framework<sup>29</sup>.

Integrating standards like SSSOM, EML, MOD, and DwC into a TIF involves careful consideration of how each standard fits within the architecture and serves the framework's goals. These standards, each with its unique focus and application domain, contribute to the richness and utility of the framework by enhancing ontology mapping, metadata management, and data interchange capabilities. In this context, the incorporation of PROV-O emerges as a strategic advantage tracing the origin and transformations of data within the TIF. With PROV-O, TIF gains the capability to provide detailed insights into the provenance of information, ensuring transparency and accountability in the processes involved. This provenance information becomes crucial for understanding how data is generated, transformed, and utilised within the TIF. Integrating DCAT is crucial for enhancing data discoverability and facilitating the sharing of metadata across different systems. In a TIF, a triplestore is an essential component for storing and querying semantic data, especially when utilising RDF for representing information. The triplestore serves as the backbone for managing structured data in the form of triples.

---

<sup>28</sup> <https://www.w3.org/2001/sw/wiki/SPARQL>

<sup>29</sup> [https://ec.europa.eu/isa2/eif\\_en/](https://ec.europa.eu/isa2/eif_en/)

## 8.1. Application Architecture

### 8.1.1. Presentation Layer

The Presentation Layer in the TIF serves as an interface for interacting with the terminological resources. It is responsible for presenting resources available at machine level through the use of APIs/SPARQL endpoints and, for rendering the terminological data and metadata in a visually intuitive manner, facilitating users access and manipulation of semantic content. Through widgets, dashboards, and interactive visualisations, it bridges the gap between complex semantic structures and end-user accessibility, promoting effective and efficient use of terminological resources within and across scientific communities.

- Modern frameworks (e.g., React, Angular) shall be adopted for developing user-friendly interfaces for data browsing, querying, and management.
- Graphic Interface tools shall be integrated for semantic data visualisation and exploration.

### 8.1.2. Application Layer

The application layer is where semantic applications and services operate, utilising the underlying semantic technologies to process, query, and manipulate semantic data. This layer is where the semantic integration and interoperability of data occurs, enabling applications to understand and utilise the meaning (semantics) of the information they process. It encompasses the tools and interfaces that allow the creation, annotation, and consumption of semantic content, aiming to make web data machine-readable and semantically rich for advanced data integration, search, and analysis across diverse domains. To enhance the framework's capability in handling ecological and environmental datasets, the use of the EML within the data integration services is emphasised. This adoption guarantees the preservation and application of comprehensive metadata, enriching the data's semantic context. Similarly, the MOD is utilised to inform users about insights present into the ontologies and vocabularies at their disposal, thereby enhancing the selection and application of semantic resources within the TIF. Furthermore, the framework uses the DwC standard for related biodiversity information, ensuring compliance with recognised standards and promoting interoperability across these specific domains. Additionally, the framework is designed to efficiently store, manage, and utilise the SSSOM files or their equivalent representations, facilitating the effective retrieval and application of ontology mappings and bolstering the framework's semantic infrastructure. Within the application layer, the triplestore is accessed via a SPARQL endpoint. The SPARQL endpoint allows users and systems to perform complex queries against the data stored in the triplestore, facilitating data exploration, integration, and analysis.

- Common and well known frameworks shall be adopted in the RESTful API development and integration with semantic web technologies.
- Adopt libraries for processing RDF data, ontology management, and SPARQL querying.

### 8.1.3. *Data storage and management*

RDF resources shall be saved in a triplestore type repository. Resources such as mappings, linked data, semantic annotation, artefact catalogue, harvesting endpoint catalogue, are examples of RDF resources. The repository triplestores shall be robust for storing RDF data and enabling SPARQL querying (e.g., Apache Jena, Stardog, GraphDB) with high performance characteristics. Metadata Repository shall contain metadata about the terminologies, including versioning information, source, and usage guidelines. Ontology and Vocabulary shall be stored in triplestores, enabling the framework to utilise complex semantic relationships and reasoning capabilities. This storage includes not only the terminologies themselves but also metadata about these terminologies, such as versioning information, provenance, and usage guidelines.

## 8.2. *Meta-Data Component*

This component is responsible for managing descriptive information about the various terminologies, datasets, and services that the framework offers.

- EML, being focused on ecological data, shall be integrated into the metadata repository. It can be used to describe datasets in terms of their ecological context, measurement standards, and metadata.
- MOD shall be directly integrated into the metadata repository, where it can be used to describe ontologies and vocabularies themselves. This includes information about ontology versions, authors, domain coverage, and other descriptive information that enhances the understanding and usability of ontologies within the framework.
- DwC, which focuses on biodiversity data standards, shall be integrated into the metadata repository and terminology database. DwC can be used to standardise the description of biodiversity data, making it an essential component for frameworks dealing with biological, ecological, or conservation-related data.
- SSSOM adoption to implement a component within the data layer dedicated to storing and managing ontology mappings in SSSOM format. This component shall be capable of handling the import, export, and update of mappings.
- PROV-O shall be used to share information about provenance of semantic resources.
- Implementing a Data Catalog Service that utilises DCAT for metadata management supports interoperability and data exchange across different systems. It enhances its compatibility with data portals and search engines, improving the discoverability of

its datasets and services. This is particularly beneficial in a federated data environment where data from multiple sources are integrated and shared.

### 8.3. Data Integration and Exchange Components

The Triplestore shall be integrated with various interoperability services in the application layer, such as data mapping, alignment services, and terminology management services. These services leverage the triplestore's capabilities to perform tasks like semantic mapping, searching for terms across different vocabularies, and dynamically linking data from disparate sources.

- Adopting RESTful APIs architecture for data access, management, and integration. REST APIs, combined with semantic web standards like JSON-LD, enable the efficient exchange and manipulation of semantic data over the web, supporting the development of interoperable web services and applications.
- Using standard RDF for data representation, RDFS and OWL for defining schemas and ontologies, and various syntax formats (Turtle, N3, XML/RDF, JSON-LD) for data interchange, the ecosystem supports a comprehensive workflow from data creation to consumption.
- Data embedding and querying can be supported by RDFa which enables the embedding of semantic data in HTML documents, enhancing the web with machine-readable data. SPARQL allows querying this data, facilitating complex data integration and retrieval tasks.
- Use SSSOM mappings to harmonise mappings from different sources, ensuring that terms from various ontologies are correctly interpreted and integrated.

### 8.4. Security and Privacy Considerations

Access control mechanisms are important strategies to avoid security problems at any level. This can involve authentication and authorisation protocols to manage access rights and permissions. For example, access control mechanisms at the triplestore level or via the SPARQL endpoint are crucial for ensuring that sensitive or restricted data is only accessible to authorised users or systems. Moreover, a control of access can help with high-power source demanding of some requests of services.

## 9. USE CASE - ITINERIS TERMINOLOGY SERVICE

The Terminology Service (TS) will be a software product dedicated to facilitating the interoperability of scientific research infrastructures through the adoption of semantic technologies. It is designed to channel and harmonise the collection of semantic artefacts, describing interdisciplinary scientific domains, produced by the infrastructures or the originating scientific community, creating a curated map (alignments and Linked Data) of relationships that capture the semantics of correspondences between terms in the artefacts.

The TS will make it possible to realise use cases that consume the map of relationships, with the effect of counteracting the proliferation of equivalent semantic artefacts, increasing semantic interoperability, and promoting the convergence on the consistent use of scientific terms. Third parties that utilise the map of relationships will have access to it through functionalities including: terminological search, artefact search, semantic annotation, search for semantically annotated informational objects, or direct access to the map to realise other use cases.

The TS will gather semantic artefacts from distributed sources and make them available to the scientific community via web services for machine-to-machine (M2M) communication and user interfaces.

For its mission to harmonise semantic artefacts of interdisciplinary scientific domains, this project entails a substantial part of the analysis of the sources of semantic artefacts, in order to create an automatic querying system, execute the mapping of terms automatically, semi-automatically, or by domain expert users. The key contributions of the terminological service to the scientific community will be two-fold:

1. Generate, curate, publish, and update the map of relationships between semantic artefacts;
2. Create services to consume the content of the map and enable new use cases.

Over time, the TS plans to increase the number of sources of semantic artefacts, which requires a growing effort on the activities of cross-monitoring the individual sources, and updating the maps of correspondences. For this reason, the TS must be designed with a high degree of automation in the workflows allowing administrative users to be advertised about required updates or maintenance operations through the detections performed by the monitoring system. This aspect is reflected in the description of each system functionality, describing the degree of automation and the desired mode of operation.

The TS, can be exposed through a dedicated web portal or integrated into pre-existing systems, primarily into the ITINERIS HUB as a first use case. Integration will be facilitated by the direct exposure of web services for M2M communication and it will also be aided by the distribution of front-end components (e.g. JS library and HTML components) that can be easily inserted into the layout of the web portal and can be styled to adhere to the look-and-feel of the host portal. The dedicated web portal will essentially consist of the combination of the same front-end elements. Front-end components will provide direct access to system functionalities, lowering any barrier to service adoption, and demonstrating the versatility of integration. Alternatively, to the use of front-end components, direct access to the web services encapsulated by the front-end components will be available.

The TS operates on informational resources represented in RDF, i.e. semantic artefacts, and for this reason, it must also be implemented using Semantic Web stack technologies

that facilitate search, accessibility, and (re)use inspired by the FAIR principles. For example, the mappings between semantic artefacts, e.g. between ontologies, between ontologies and thesauri, or between thesauri, must be represented in the most suitable semantics. Semantic associations between terms must be represented in RDF and query services to investigate the content of semantic artefacts must be exposed through SPARQL endpoints. In addition, metadata of semantic artefacts and their mappings must comply with the standards used by the community (e.g., DCAT, SSSoM, MOD v. 2.0). Front-end components of the TS must be created and distributed using libraries that employ standard, up-to-date web programming languages, justified by the functionalities to be implemented (e.g., JS library and HTML components). It is essential to seek and ensure compatibility with the most widely used browsers (Google Chrome, Microsoft Edge, Mozilla Firefox, Apple Safari), and properties of adaptability and responsiveness to the main desktop, tablet, and mobile resolutions.

### 9.1. Terminology Service Requirements

Below are listed the main system features required for the proper functioning of the TS and defined in the published invitation to tender.

**Integrability:** The TS will be integrable in two ways: through a set of web services that expose the functionalities of the system, and through front-end components (e.g., JS library and HTML components) that encapsulate these services, offering visual interaction elements. Research infrastructures, or more generally host sites, will be able to develop custom front-end elements that invoke the system's web services, or integrate the front-end elements (e.g. search bar, filters, facets, results area), with the option to style them according to the look-and-feel of the host portal. In some instances (such as annotation), integrability must ensure the bidirectionality of data transmission between the TS and the existing data and metadata management computer systems (e.g. service catalogues).

**Scalability:** The back-end of the TS must allow deployment on a server with dynamic loading of hardware resources; at the same time, it must ensure options for configurability of hardware resources associated at boot time (e.g. RAM size), and system variables for services and interaction protocols. The goal is to be able to cope with service stress situations when an initial calculation of concurrently active users cannot be made. The front-end must provide solutions for a flawless experience, in case it is necessary to display semantic artefacts with a very high number of terms (100+), or the display of numerous nodes and connections (50+) in the networks of relationships between terms.

**Internationalisation:** The front-end of the TS will need to include a configurable system to change the user interface language, and to consequently add new translation languages (e.g. with a translation file for the labels of graphic components).

**Modularity:** The TS must be designed in a modular and non-monolithic manner, through a design based on microservices, with the goal of creating minimal dependencies between the components that perform the functionalities of: *i.* managing the lifecycle of each component, *ii.* integrating new components easily, or *iii.* updating existing components without negatively impacting the entire architecture. The TS must ensure that in the event of a service/component failure, the other services/components continue to function end-to-end. The TS must be designed with a client-server infrastructure model.

**Automation:** Over time, the TS anticipates an increase in the number of sources of semantic artefacts, which necessitates a growing effort on the activities of cross-monitoring the individual sources, and maintenance of the maps of correspondences. For this reason, the TS must be designed with a high degree of automation in workflows allowing administrative users to be advertised about required updates or maintenance operations through the detections performed by the monitoring system. This aspect is reflected in the description of each system functionality, describing the degree of automation and the desired mode of operation.

## 9.2. Terminology Service Functionalities

### 9.2.1. Collection of Semantic Artefact Sources

The collection of semantic artefact sources can be carried out by authorised administrative users of the TS, or it can be requested by users from research infrastructure who submit their collection of semantic artefacts through an application process so that it becomes part of the map constructed by the TS. A source of semantic artefacts (e.g. <https://ecoportal.lifewatch.eu/>) is registered and subsequently subjected to harvesting. The registration of a source occurs through a form that collects: *i.* attributes described with standard vocabularies (e.g. PROV-O, DCAT, etc.), taking into account, for instance, the source's provenance (affiliation); *ii.* a list of domains covered by the artefacts; *iii.* the address of the endpoint exposing the services for harvesting (API or SPARQL); *iv.* a link to the source code repository of the harvesting interface implementation; *v.* a Single Point of Contact (SPOC) for future communications. At the end of the submission, the system processing the request performs a test on the harvesting interface to verify the successful connection and retrieval of the necessary information. The test result is recorded with the source and returned to the user (proposing party) with a feedback message; an alert is also sent to the administrative user. All submissions made through the form remain in pending review and approval by an administrator. Upon approval, the source will officially be added in the catalogue and it will be subjected to the scanning and discovery of semantic artefacts included therein (harvesting). At this point, the semantic artefacts will become part of the catalogue and subsequently used for terminological collection.

The list of sources, or "Semantic Artefact Catalogue", is a resource actively maintained by the system's administrative users. Only these users will have access to an administration

dashboard that facilitates management processes. For example, an administrative user can view the following attributes for each item in the catalogue: *i.* the source's provenance (affiliation); *ii.* a list of domains covered by the artefacts; *iii.* the date of the last harvesting; *iv.* the monitoring status of the endpoint (e.g. Active, Inactive); *v.* feedback on the status of the last harvesting (e.g. Success, Failed, Error, etc.); *vi.* the address of the source endpoint; *vii.* a link to the source code repository of the harvesting interface implementation; *viii.* the contact of a SPOC for reporting any issues; *ix.* a form to send emails to the SPOC and so on. From this page, the administrator can view the health status of all sources, the state of harvesting processes, open details on the harvesting process for each source, notify the responsible user in case of issues detected in the harvesting process, or communicate directly to the SPOC the type of problem encountered. They can enable or disable a source for inclusion in the harvesting process; enable an unscheduled harvesting process; and additional functions to be defined during the development and design phases.

### 9.2.2. *Harvesting*

Harvesting is a functionality of the TS that consists of two separate but dependent subprocesses: collection of semantic artefacts and terminological collection.

#### 9.2.2.1. *Collection of Semantic Artefacts*

The collection of semantic artefacts is the set of automatic operations necessary to discover the (new) content of different sources of semantic artefacts for their integration within the catalogue. The collection will operate with a process that at regular frequency: *i.* scans the source catalogue; *ii.* performs harvesting and updates on the semantic artefact catalogue; *iii.* records the monitoring status of the source; *iv.* informs the TS administrator of any process failures, who in turn can configure the sending of manual or automatic messages to the SPOC about the failures (e.g. based on process harvesting log monitoring).

#### 9.2.2.2. *Terminological Collection*

The terminological collection is the automatic process that allows investigating the content of the semantic artefacts present in the semantic artefact catalogue. The terminological collection, together with the collection of semantic artefacts, must produce all the data necessary to support mappings (terminological alignment) and networking operations, and other TS functionalities that depend on the harvesting.

### 9.2.3. *Terminological Alignment (Lexicographic Mapping)*

The mapping of terms between semantic artefacts can be carried out manually or in a supervised manner, with the support of a textual analysis component of terms. In the first case, an expert user will have a front-end component (e.g. JS library and HTML

components) to load two semantic artefacts (pointing to their repository), view them, and manually indicate correspondences between the two, choosing the most appropriate semantic relationship from the ontology of relations. In the second case, the expert user can task a textual analysis component of terms (e.g. label similarity algorithms) to analyse and display suggested alignment results, with the goal of accepting or rejecting the proposed alignment. In both cases, the user can save the mapping result and its metadata (in RDF and other formats), which will be added to the existing map.

It will be possible to perform maintenance of the maps in order to respond to eventual updates of the involved semantic artefacts. Update events can include the deprecation or deletion of a term that is subject or object of an alignment, as well as any other different element that the TS can detect using the diffing functionality. If the diffing detects that there are new terms, for example, in a semantic artefact, then it can notify the expert to update the mapping. At the end of a map update, the administrative user saves the latest version, and the system archives the previous version. Since mappings are performed between semantic artefacts, the TS must be able to evaluate the transitivity and symmetry of alignment relations, expand the mapping, and use it in term recommendation functions, such as terminological search, or semantic annotation, or the search for semantically annotated informational resources.

#### *9.2.4. Ontology of Mapping and Linked Data*

The semantic artefacts hosted in research infrastructures are heterogeneous in type and in the metamodel used to represent them. A mapping of similarity between terms must foresee the non-trivial case of mapping different types of artefacts. The sought effect will be such that, given a term, it will be possible to retrieve the terminological alignments generated between all semantic artefacts either if they are built with the same metamodel or not. The Linked Data generated through the semantic artefacts in research infrastructures must respond to "design patterns" highlighted in each scientific domain (e.g. Ecology, Biodiversity, Oceanography, Environmental Biology, Biodiversity Conservation, Earth Sciences). In this way, networking operations (Linked Data) will be based on a consensus of the scientific community that has approved the patterns. The collection of terminological alignment relations and the collection of design patterns for Linked Data must be created and maintained in a dedicated ontology to capture these aspects of scientific domain and support to the TS, with the primary effect of harmonising the exchange of information with service users.

#### *9.2.5. Networking (Linked Data)*

Networking of terms between a pair of semantic artefacts can be carried out manually by an expert user who will have access to a front-end component (e.g. JS library and HTML components) to load two semantic artefacts (pointing to their repositories), view them, and manually identify correspondences between the two, selecting the most appropriate

semantic relationship from the ontology of relationships. The user can save the networking result (in RDF and other formats), which will be integrated into the existing network (linked data). The front-end components must be styleable (CSS) to adhere to the look-and-feel of the host portal.

It will be possible to perform maintenance of the network in order to respond to updates of the involved semantic artefacts. Update events can include the deprecation or deletion of a resource that is a subject or object in a network, as well as any other element of difference that the TS can detect using the diffing functionality. If the diffing system detects that there are new terms in a catalogue of artefact, it can notify the expert to update the network of relationships. At the end of a network update, the user saves the latest version, and the system archives the previous version.

#### 9.2.6. *Diffing*

Diffing between two versions of the same semantic artefact is a functionality enabled on the semantic artefact catalogue, which, at regular intervals or through direct activation by the administrative user, performs tests on sources to detect changes. Diffing must also detect and keep track of the presence of a semantic artefact in multiple sources possibly at different stages of development. In this case, the TS saves this information, for example, by explicitly relating identical artefacts, their provenance, and any detected version (map of artefacts in the semantic artefact catalogue). To initiate the diffing on the semantic artefact catalogue, an administrative user will have a control panel available with the option to launch the diffing across the entire source catalogue or individually on one of the sources. The result of the diffing is displayed on a dashboard where the diffing status of semantic artefacts in the source will be clearly visible, with a reference to the diffing result, and the list of alignments and networks affected by the detected changes. The diffing function supports the maintenance of alignments and networks, which in turn affect the outcome of all functions that consume the maps and linked data. The administrative user can update the maps corresponding to the semantic artefacts that have been updated.

#### 9.2.7. *Semantic Annotation*

Semantic annotation is an application function of terminological search, to associate terms in search results with informational objects such as articles, datasets, web services, etc. The Semantic Annotator presents the terms from the Terminological Search, suggested based on keywords entered by the user, and ranked according to ranking criteria. The terms in the results are enriched with metadata of provenance, scientific domain, and resource type for further filtering and selection. The semantic annotation function will be distributed in two modes: through a front-end component (e.g. JS library and HTML components), and through a web service for integration into research portals.

In the first mode, the user interacts with a registration form for the informational object, with a dedicated space for annotation. The user can fill out the form with the necessary data (e.g. title, description, author, conference, scientific domain, and other inputs required according to a predefined RDF metadata schema) and submit the form. The form submission provides the TS with the form data content to be entered into its register of annotated objects.

In the second mode, the annotation function must be exposed through a web service, which similarly allows saving the object and the annotation metadata, coming from the service consumer.

For both modes, the annotation functionality must ensure the retention in the TS of data on the annotated object and the terms used for annotation. These data must be saved in the TS system for use in searching for annotated objects and navigating the graph of annotations.

#### *9.2.8. Storage of Mappings and Linked Data*

Any type of RDF resource generated by the use of the TS functionalities will be saved in a triplestore type repository. Resources such as mappings, linked data, semantic annotations, semantic artefact catalogues, harvesting endpoint catalogues, are examples of RDF resources. The repository must be a software product distributed for free or with a community licence (e.g. Ontotext, Apache Jena, RDF4J, Blazegraph, etc.), must have high performance characteristics, and compatibility with the SPARQL language in its latest version, and functionalities that would bring additional value to the TS (e.g. read and write access control, timeout management, support for quad-store, extension of the SPARQL protocol for textual search, etc.).

#### *9.2.9. Search for Semantic Artefacts*

This search function allows entering keywords and receiving a list of semantic artefacts, enriched with metadata, from the semantic artefact catalogue. The mode of textual search will be simple, facilitated (e.g. autocomplete) and advanced (e.g. use of wildcard and partial match, etc.). The keyword search can be performed on a range of textual elements: local names, URIs, labels, with the ability to enable/disable each of these options. Textual search can be performed on the metadata of the entire semantic artefact with the ability to filter metadata types. The search result can be filtered through the use of facets on the metadata associated with the semantic artefacts (e.g. provenance-affiliation, type of artefact, date of publication, scientific topic, available languages, authors, vocabularies used, etc.).

To perform the search, the user (human or machine) will have a front-end component (e.g. JS library and HTML components) to enter keywords and operate on the customisation

elements of the search, or a web service capable of collecting the same request according to defined parameters. To display the search result, a user will have a front-end component available, with a clear presentation of the results and metadata, and elements for faceting the results, or a JSON data structure in case of invoking the web search service. The front-end components must be styleable (CSS) to adhere to the look-and-feel of the host portal.

#### *9.2.10. Access to Semantic Artefacts*

Individual semantic artefacts, originating from distributed sources, will be navigable through *ad hoc* developed widgets with varying levels of complexity (e.g. hierarchy). Another form of access to semantic artefacts is the SPARQL query protocol. The TS will provide a front-end component for formulating, validating, and executing SPARQL queries directed at the semantic artefact's source repository.

#### *9.2.11. Terminological Search*

The difference between the search for semantic artefacts and terminological search is that the latter returns a list of terms while the former returns a list of semantic artefacts containing the terms. This search function allows users to input keywords and receive a list of terms with associated metadata, sourced from the semantic artefact catalogue. The textual search mode will be both simple and advanced (e.g. use of wildcard and partial match, etc.). Keyword searches can be conducted across a range of textual elements: local names, URIs, labels, with the ability to enable/disable each of these options. The search result can be filtered through the use of facets on the metadata associated with the semantic artefacts containing the term (e.g. source-affiliation, type of artefact, publication date, scientific topic, available languages, authors, vocabularies used, number of mappings by type, etc.). To conduct the search, users will have access to a front-end component (e.g. JS library and HTML components) to input keywords and manipulate the search customisation elements, or a web service capable of processing the same request according to defined parameters. To display the search result, a user will have a front-end graphical element, with a clear presentation of the results and metadata, and elements for faceting the results. In the case of invoking the web search service, the result set will be described in a JSON data structure.

#### *9.2.12. Ranking of Terms in the Terminological Search Results Set*

The terminological search process will operate with a ranking system to support the selection of terms in the search result. The ranking criteria, to be defined during the development phases, might be based on available metadata: the latest update of the term, source, availability of multilingualism, etc. Other criteria may derive from the analysis of the semantic surroundings of the term (based on the network of semantic relationships).

For example, in the case of terminological search for semantic annotation, the ranking could order terms based on the number of mappings with terms from other semantic artefacts, whether the containing artefact is from the same scientific domain as the informational object to be tagged, and whether the term has a dense network of relationships with semantic artefacts from interdisciplinary domains. An informational object (e.g. dataset) annotated with a high-ranking term is more likely to be directly found in proportion to the popularity of its source, the mapped sources, and the interdisciplinary domains. Other ranking elements could come from user community feedback, for example, the frequency of use of the term. Although primarily supporting terminological search, this functionality could be isolated and used to calculate relevance scoring on semantic artefacts in a collection or a single external artefact submitted for evaluation.

#### *9.2.13. Search for Semantically Annotated Informational Objects*

The TS enables the search for informational objects on portals using the semantic annotation function. This functionality is available because the TS has recorded the annotated data on which the search is performed. The search is possible through the use of keywords, through the navigation of semantic artefacts, and also through the visualisation and navigation of the graph of connections between terms and semantically annotated objects (annotations graph). In the first case, the user has a textual field to enter keywords. In the second case, the user can navigate the structure of a semantic artefact and, depending on the selected term, receive the list of annotated objects. In the third case, the user can visualise the annotations graph starting from a term or a semantically annotated object and navigate the graph by expanding or reducing nodes and relationships. The details of the semantically annotated object shall include a link to the object hosted in its source. This search function is distributed through a front-end component (e.g. JS library and HTML components) that can be integrated into an existing portal, providing the search field, navigation search functionality, and results display area. The front-end components must be styleable (CSS) to adhere to the look-and-feel of the host portal. This search functionality is also exposed as a web service to enable integration into more complex processes.

#### *9.2.14. Dereferencing, Publishing, Editing*

To make the user experience as smooth as possible and encourage the scientific community at large to publish their semantic artefacts in a repository/registry that can enhance their FAIRness, it will be necessary to enhance one of the sources in some of its components. The enhancement will concern the EcoPortal repository, whose long-term management and sustainability are guaranteed by one of the infrastructures involved in the project. During the analysis and design phase, the components to be enhanced will be detailed to include updates on the source code to the latest available version, development and integration of the dereferencing functionality for URIs, visualisation of languages on the annotation

properties of semantic artefacts, updates on the editing tool (<https://vocbench.uniroma2.it/doc/>) and its integration into EcoPortal. Since the enhancement involves modifying the code at all levels, it is necessary that this occurs in close contact and in full synergy (or coordination) with the developers, system administrators, and generally the administrators of the repository subject to enhancement.

Furthermore, the dedicated portal of the TS must have a simplified and automated loading and application system that allows the request for loading and publishing the resource on EcoPortal. This integration will be structured in a smooth workflow that allows, with maximum automation, to load and apply for the publication of the semantic artefact on EcoPortal, and subsequently, upon administrator approval, subject to harvesting in the TS semantic artefact catalogue.

#### *9.2.15. Creation and Management of Users and Roles*

Based on the functionalities described above, users will be able to use the TS anonymously or authenticated with the assignment of roles and privileges. The user account creation system must comply with the European GDPR regulation on personal data processing. The login system must use existing authentication protocols (e.g. Single Sign-On). Similarly, the use of constrained work sessions and protection against service request overload must be provided. These and any other security measures, to be specified during the project development phase, aim to eliminate the risk of digital impersonation and protect the system from unauthorised use of functionalities among administrative users and TS consumers. A user will be able to delete their account, and the TS must ensure that the contributions (e.g. mappings) provided by the user are preserved in the system while having the ability to delete personal data considered sensitive (e.g. first name, last name, email).

## 10. CONCLUSION

The ITINERIS project marks a significant evolution in the field of environmental research within Italy, by introducing a Terminology Interoperability Framework (TIF) aimed at overcoming the challenges posed by data heterogeneity and fostering a collaborative research ecosystem. Through the compilation of an extensive array of terminologies, the document highlights the crucial importance of a unified semantic framework to facilitate communication across diverse research domains.

At the core of this endeavour, principles such as Integrability, Scalability, Internationalisation, Modularity, Automation, Reusability, and User Centricity guide the development of advanced semantic technologies. These principles underscore adaptability, inclusivity, and efficiency, essential in the face of evolving scientific inquiries and technological advancements.

The adoption of international standards underscores a commitment towards the reliability and integration of the TIF with global research initiatives, extending the impact of the ITINERIS project beyond national borders. Furthermore, the emphasis on user-centred design and accessibility within the TIF demonstrates the project's dedication to creating an inclusive and collaborative research environment, enhancing the efficiency of research processes and democratising access to scientific data.

The TIF addresses not only the immediate needs of the environmental research community but also lays the groundwork for a more interconnected and collaborative scientific ecosystem, encouraging a collective commitment towards FAIR data and a more integrated and accessible scientific research landscape.

## 11. REFERENCES

A. Chatterjee, N. Pahari, and A. Prinz, “HL7 FHIR with SNOMED-CT to Achieve Semantic and Structural Interoperability in Personal Health Data: A Proof-of-Concept Study.,” *Sensors (Basel)*, vol. 22, no. 10, May 2022, doi: 10.3390/s22103756.

IEEE Std 2030™, IEEE Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation with the Electric Power System (EPS), End-Use Applications, and Loads.

T. Ford, J. Colombi, S. Graham, and D. Jacques, “Survey on Interoperability Measurement,” p. 67, Jun. 2007.

P. Gottschalk, “Maturity levels for interoperability in digital government,” *Gov. Inf. Q.*, vol. 26, no. 1, pp. 75–81, 2009, doi: <https://doi.org/10.1016/j.giq.2008.03.003>.

T. Jepsen, S. Mithas, C. Hsu and G. Kraft. “Healthcare IT,” *IT Prof.*, vol. 12, no. 02, pp. 14–16, 2010, doi: 10.1109/MITP.2010.57.

M. B. Jones, M. O’Brien, B. Mecum, C.Boettiger, M. Schildhauer, M. Maier, T. Whiteaker, S. Earl, S. Chong, 2019. “Ecological Metadata Language version 2.2.0,” doi: 10.5063/f11834t2.

N. Karam, C. Müller-Birn, M. Gleisberg, D. Fichtmüller, R. Tolksdorf, and A. Güntsch, “A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data,” *Datenbank-Spektrum*, vol. 16, no. 3, pp. 195–205, 2016, doi: 10.1007/s13222-016-0231-8.

Y. Le Franc, L. Bonino, H. Koivula, J. P. Essen, and R. Pergl, “D2.8 FAIR Semantics Recommendations Third Iteration.” Zenodo, 2022, doi: 10.5281/zenodo.6675295.

R. Rezaei, T. K. Chiew, S. P. Lee, and Z. Shams Aliee, “Interoperability evaluation models: A systematic review,” *Comput. Ind.*, vol. 65, no. 1, pp. 1–23, 2014, doi: <https://doi.org/10.1016/j.compind.2013.09.001>.