Metadata, Controlled Vocabularies & Ontologies 1st Working Meeting Report

Lecce, 12-13 November 2018

Authors

Caterina Bergami, LifeWatch Italy (CNR ISMAR)
Nicola Fiore, LifeWatch ERIC
Alessandro Oggioni, LifeWatch Italy (CNR – IREA)
Ilaria Rosati, LifeWatch Italy (CNR – IRET)

With the contribution of

Daphnis De Pooter, LifeWatch Belgium
Nikos Minadakis, LifeWatch Greece
Gregor Aljancic, LifeWatch Slovenia
Magda Aljancic, LifeWatch Slovenia
Paul Martin, LifeWatch Netherlands
Liao Xiaofeng, LifeWatch Netherlands
José María García, LifeWatch Spain
Antonio José Sáenz-Albanés, LifeWatch ERIC
Cristiano Fugazza, LifeWatch Italy (CNR- IREA)

www.lifewatch.eu

Executive Summary

On November 12th & 13th, the LifeWatch-ERIC (hereafter LW-ERIC) Service Centre organised the first working meeting on "*Metadata, Controlled Vocabularies and Ontologies*" in Lecce, Italy.

The aim of the meeting was to set a roadmap for a common strategy to be adopted on metadata, controlled vocabularies and ontologies within the LW-ERIC community and in accordance with the FAIR principles.

The meeting involved participation from 13 experts, with both scientific and technical backgrounds, from six national nodes of LW-ERIC (Belgium, Greece, Italy, Spain, Slovenia, The Netherlands). Three sessions were held on 1) Metadata, 2) Controlled Vocabularies, and 3) Ontologies, each one including a tour de table on existing approaches in the national nodes and a discussion to define common best practices for the implementation and curation of metadata, standardized controlled vocabularies and ontologies.

Presentations, foreseen for each session, outlined previous work and set the current landscape of models, tools and technologies that are available to support metadata, controlled vocabularies and ontologies within national nodes of LW-ERIC. They also provided material for following discussions, where the main technical and scientific approaches about the semantic issues in the Research Infrastructures have been also considered in order to identifying a common strategy to be adopted by LW-ERIC.

This meeting is a first step toward medium-term goals. Follow-on work, including more meetings or workshops by experts and the LW-ERIC community, will significantly advance this initiative.

Table of Contents

Metadata

Introduction

Metadata is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or description."

Metadata (hereafter MD) summarizes basic information about data, which can make finding and working with particular instances of data easier. For example, author, date created and date modified and file size are examples of very basic document metadata. Having the ability to filter through that MD makes it much easier for someone to locate a specific document.

In addition to document files, MD is used for images, videos, spreadsheets and web pages. The use of MD on web pages can be very important. MD for web pages contain descriptions of the page's contents, as well as keywords linked to the content. These are usually expressed in the form of metatags. The MD containing the web page's description and summary is often displayed in search results by search engines, making its accuracy and details very important since it can determine whether a user decides to visit the site or not. Metatags are often evaluated by search engines to help decide a web page's relevance, and were used as the key factor in determining position in a search until the late 1990s. The increase in search engine optimization (SEO) towards the end of the 1990s led to many websites "keyword stuffing" their MD to trick search engines, making their websites seem more relevant than others. Since then search engines have reduced their reliance on metatags, though they are still factored in when indexing pages. Many search engines also try to halt web pages' ability to thwart their system by regularly changing their criteria for rankings, with Google being notorious for frequently changing their highly-undisclosed ranking algorithms.

MD can be created manually, or by automated information processing. Manual creation tends to be more accurate, allowing the user to input any information they feel is relevant or needed to help describe the file. Automated MD creation can be much more elementary, usually only displaying information such as file size, file extension, when the file was created and who created the file.

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*. The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data. The first step in (re)using data is to find them. MD and data should be easy to find for both humans and computers. Machine-readable MD are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process. First principle of FAIR is "**F1. (Meta)data are assigned a globally unique and persistent identifier**" dedicated of assign globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of MD. Within LW-ERIC we propose to have a MD information of different entities of environmental research aspects in order to improve the discovery of data.

Tour de Table Presentations

The session about MD began with a presentation by Alessandro Oggioni with some questions about the new need about MD from LW-ERIC. In particular he asks: Which type of entities do we want to describe? Only the dataset? Does LW-ERIC manage entities as instruments, persons, programmes/projects, activities, etc.? Which MD standard? Which MD profile? For which entities is it important provide a PID (URI!)?

The session continues with presentations by LifeWatch Spain, Greece, LifeWatch Netherlands, LifeWatch Belgium and LifeWatch Italy about the current practices within each node. The discussion continues with the intervention of Antonio José Sáenz that describes the scenario internally at the LW-ERIC. Within LW-ERIC there are several infrastructures, projects that has adopted at national level different standard and schemas. LW-ERIC needs to integrates all this contributions and to deploy it as soon as possible.

Finally Antonio José Sáenz, according with Alessandro Oggioni original idea, proposes to collect a list of entities, a list of MD schema/standard of this entities, and a list of tools (spreadsheet form).

Discussion

After the tour de table is proposed to collect a list of entities that LifeWatch wants to consider, a list of MD schema/standard of different entities, and a list of tools to manage different MD.

We provided 2 spreadsheets for “Metadata entities” and “Metadata tools” (spreadsheet form) in order to allow to collect, from all the participants, proposals of entities to be metadata, schemes and tools to carry out the MD.

The tables below summarize what was collected by the various participants about entities and schema/standard (Table 1) and tools/software for editing, curate, storage, sharing, harvesting, querying and mapping (Table 2).

Table 1. Entities and their definition, MD schema/standard and link to MD Schema documentation.

Entities	Definition of entities	MD schema/standard	MD Schema link documentation
Dataset	Ecological Metadata Language (EML) is a metadata specification particularly developed for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts.	Ecological Metadata Language (EML), ISO19115 Parthenos Entity Model	http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language http://www.parthenos-project.eu/

Entities	Definition of entities	MD schema/standard	MD Schema link documentation
Network	Administrative or organisational grouping of Environmental Monitoring Facilities managed the same way for a specific purpose, targeting a specific area.	Environmental Monitoring Network (EMN)	https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf
Site	A georeferenced object directly collecting or processing data about objects whose properties (e.g. physical, chemical, biological or other aspects of environmental conditions) are repeatedly observed or measured.	Environmental Monitoring Facility (EMF), CRM-Geo	https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf http://www.cidoc-crm.org/
Station	A georeferenced object directly collecting or processing data about objects whose properties (e.g. physical, chemical, biological or other aspects of environmental conditions) are repeatedly observed or measured.	Environmental Monitoring Facility (EMF)	https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf
Instrument	Any type of sensors or processes that observed specific phenomenon	SensorML	https://www.opengeospatial.org/standards/sensorml
Scientific Name	The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic rank that can be determined.	Dublin Core (DC) and Darwin Core (DWC), CRM Sci, MarineTLO	http://dublincore.org and https://dwc.tdwg.org http://www.cidoc-crm.org/
Publication	Paper, book or any type of documents	Common European Research Information Format (CERIF)	https://www.eurocris.org/cerif/main-features-cerif
People	Names, roles and contact details of people involved in research	Friend Of A Friend (FOAF), CIDOC-CRM	http://www.foaf-project.org
Activity	Specific set of AbstractMonitoringFeatures used for a given domain in a coherent and concise timeframe, area and purpose.	Environmental Monitoring Activity (EMA), CIDOC-CRM, CRM-Sci	https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf http://www.cidoc-crm.org/

Entities	Definition of entities	MD schema/standard	MD Schema link documentation
Programme Project	Framework based on policy relevant documents defining the target of a collection of observations and/or the deployment of AbstractMonitoringFeatures on the field.	Environmental Monitoring Programme (EMP), Parthenos Entity Model	https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf
Sample	IGSN stands for International Geo Sample Number. The IGSN is an alphanumeric code that uniquely identifies samples from our natural environment and related sampling features.	International Geo Sample Number (IGSN) CRM-Sci	http://www.geosamples.org/igsnabout http://www.cidoc-crm.org/

Table 2. Tools to manage the different MetaData.

Steps ->	Editing	Curation	Storage	Sharing	Harvesting	Querying	Mapping	*Transformation*
Entities						metaphactory	3M	
Dataset	for spatial data and for describe it using ISO19115 - EDI client metadata editor			GeoNetwork, pyCSW, Morpho, ...	GeoNetwork, pyCSW, Morpho, ...	metaphactory, GeoNetwork, pyCSW, Morpho, ...	3M	x3ml Engine
Network	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Site	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Station	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Instrument	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Specie						metaphactory	3M	x3ml Engine
Publication						metaphactory	3M	x3ml Engine

Steps ->	Editing	Curation	Storage	Sharing	Harvesting	Querying	Mapping	*Transformation*
People	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Activity	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Programme/Project	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Sample	EDI client metadata editor	EDI server metadata editor	EDI server metadata editor			metaphactory	3M	x3ml Engine
Infrastructure						metaphactory	3M	x3ml Engine

Table 2 will have to be further discussed to get a list of tools and software that allows users to achieve different steps. To better clarify what should be included in the table can be proposed 2 examples: ElasticSearch is a software for storage entities in RDF or XML format and also for querying, but is not really software for end users; GeoNetwork is a software for store and share MD about dataset but it is not for share MD about other entities.

The participants agree to introduce these approaches to the Executive Board and a series of working meetings (face2face and virtual) will be planned for the next year to continue the work.

Controlled Vocabularies

Introduction

A controlled vocabulary is an organised arrangement of words (concepts) used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and alternative terms and has a defined scope or describes a specific domain. Controlled vocabularies provide a definition, a coding scheme and globally unique and persistent identifier for each term. The level of detail of controlled vocabularies ranges from short unidimensional lists to complex vocabularies with hierarchical relationships (Thesauri).

Controlled vocabularies play an important role in metadata standards, because they define the meaning of metadata elements and the values allowed in an element/attribute. Apart from that, they can also help to find relevant data, or provide information on how to interpret data (for both, humans and machines) and reuse it. The use of controlled vocabularies helps to improve the interoperability of data, as vocabularies facilitate the interpretation and harmonization of data (especially, if other researchers employ the same vocabulary for their data). When research communities agree to use common language for the concepts in metadata and data, then the discovery, linking, understanding and reuse of research data are improved.

In addition to selecting metadata standards or schemas, controlled vocabularies will also form a crucial part of the LW-ERIC infrastructure. They will be used in metadata annotation and for labelling data files in order to make information findable, accessible, interoperable, and re-usable (FAIR data principles).

Therefore LW-ERIC needs to define a common strategy for the implementation and curation of new standardized terminological resources but also for the improvement and alignment of those already produced by the different national nodes.

Tour de Table Presentations

The session began with a tour de table with short presentations by national nodes working and/or using controlled vocabularies. The aim was to provide an overview of the current practices for the implementation and management of controlled vocabularies within each node and to get the background for following discussions.

LifeWatch-Belgium presented a list of vocabularies currently used for metadata enrichment and data annotation in their Dataset Catalogue. In particular they use existing terminological resources such as the Darwin Core glossary of terms; the Eunis Habitat classification; the Marine Regions standard list related to marine georeferenced place names and areas; the World Register of Marine Species (WoRMS) which provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy; and the BODC controlled vocabularies (P01: BODC Parameter Usage Vocabulary; P06: BODC-approved data storage units; Q01: OBIS sampling instruments and methods attributes; L22: SeaVoX Device Catalogue; S10: BODC Biological entity gender; S11: BODC Biological entity development stage; M20: Marine Habitat Classification for Britain and Ireland).

Moreover, LifeWatch-Belgium also produced the Marine Species Traits Vocabulary (http://www.marinespecies.org/traits/wiki), a hierarchical list of traits, built by a customised version of the open source Semantic MediaWiki (SMW), which was established within the VLIZ hosted Coastal Wiki (http://www.coastalwiki.org). The hierarchy of traits has been organised into a series of four "collections": Biological Descriptors, Distribution Descriptors, Ecological Descriptors and Species' Importance to Society for a total of 688 concepts.

LifeWatch Belgium highlighted some issues related to the use of Semantic MediaWiki for implementing of controlled vocabularies. In particular they experienced issues with:

1. Bugs inherent to the system (some relations between the concepts didn't always show up even though they should);
2. Bugs which are introduced with each update of media wiki (which also means you can't just copy paste the code from the Coastal Wiki or the TDWG wiki because there will be some compatibility issues).

Moreover, editors actually didn't use the SMW to develop the content of the vocabularies (people just preferred to exchange excel sheets), it was only used to share the content online after the content was agreed upon.

LifeWatch-Italy presented the approach followed for the development of thesauri on functional traits of several groups of aquatic organism (Phytoplankton, Macrozoobenthos, Zooplankton, Fish and Macroalgae) and also thesauri on alien species, endemism, genomic and barcoding. They are all available online and some already published on the services catalogue of the LW-ERIC service centre (http://www.servicecentrelifewatch.eu/catalogue-of-services). The LW-ITA thesauri are concept schemes including a set of concepts identified unambiguously by a URI and labelled with a term, which is defined by one or more lexical strings. Terms are selected from natural language and each term is used to represent only one concept. They are defined through the SKOS format (Simple Knowledge Organization System) which provides a standard way to represent knowledge organization systems using the Resource Description Framework (RDF). Actually, the LW-ITA thesauri are edited in English and contain the following information for each concept:

- Uniform Resource Identifier (URI);
- Preferred term/Preferred (*skos:prefLabel*);
- Non preferred term/Alternative label (*skos:altLabel*);
- Notes (*skos:definition; skos:note; skos:scopeNote; skos:historyNote);*
- Semantic Relationships (hierarchical relationships: *skos:broader* and *skos:narrower;* linking relationships: *skos:narrowMatch, skos:broadMatch, skos:closeMatch, skos:relatedMatch, skos:exactMatch).*

In order to produce SKOS formatted thesauri, LW-ITA used TemaTres, an open source and web-based tool. TemaTres includes, a simple but functional user interface for editing concepts, sophisticated search capabilities and other functionalities such as:

1. No limits to number of terms, alternative labels, levels of hierarchy, etc…;

2. Import from Skos-Core, tabulated or tagged text file;
3. Multilingualism;
4. Relationships between terms;
5. Notes (for editor and user);
6. Options for linking to external concepts;
7. Export in a number of standardized forms (e.g. Mads, Skos, Vdex, XTM, Json);
8. User management;
9. Advanced Reports for editors in CSV;
10. SPARQL endpoint and API.

Moreover LW-ITA used Silk (the Linked Data Integration Framework) an open source framework for the discovery of links among RDF resources, for the mapping and the alignment of LW-ITA thesauri. SILK is based on the Linked Data paradigm, which is built on two simple ideas: first, RDF provides an expressive data model for representing structured information; second, RDF links are set between entities in different data sources. Using the declarative *Silk - Link Specification Language* (Silk-LSL), developers can specify which types of RDF links should be discovered between data sources, as well as which conditions data items must fulfil in order to be interlinked. Silk accesses the data sources that should be interlinked via the SPARQL protocol, and can thus be used against local, as well as remote, SPARQL endpoints. Link Specifications can be created using the Silk Workbench graphical user interface, which guides the user through the process of interlinking different data sources.

In order to find out a complete connection among LW-ITA thesauri, a two-step process has been put in place: first, SILK has been applied by comparing preferred terms from two separate thesauri using the Token-wise distance algorithm to discover new links; then the SILK results have been validated by editors in order to verify the accuracy of the links and identify the most suitable types of interlinking property (i.e., *skos:exactMatch* or *skos:closeMatch*).

Lifewatch-Italy also presented the editorial organization illustrating the working groups, the roles and the implementation workflow. The process of LW-ITA thesauri implementation is a collaborative process involving different working groups with specific roles: editors, ICT experts and validators. Editors are experts of the specific knowledge domain and they have the responsibility of each aspect of the thesaurus construction and management, from planning to design, dissemination and maintenance. The ICT group supervises the technological aspects of thesauri modelling, advising on semantic technology and modelling, and giving technical support to the editor team in the selection, use and maintenance of the most suitable tools for the development of thesauri and their linking. They collaborate with the editor team for defining relationships between concepts and data type properties for defining attributes (or qualities) of concepts. Validators are domain experts who review the constructed thesaurus and highlight any question about the terms chosen, any gap, missing or redundant feature, as well as any usability issue.

The implementation workflow envisages four phases: (i) terms research and selection, (ii) formalisation, (iii) edition, and (iv) validation of the thesaurus. More information are available in the *LifeWatch Italy Thesauri Documentation, Version 1.0 (***http://www.servicecentrelifewatch.eu/web/lifewatch-italia/publications***).*

Finally, LW-Netherlands shared its experience in the creation of the ENVRI reference model and it could make available the vocabulary at now used for describing 'things found in RIs' (activities, services, actors, facilities, etc.) in the OIL-E ontology http://oil-e.net/ontology/.

Discussion

After the tour de table, the main points of discussion were:

1. Tool for the management of controlled vocabularies in LW-ERIC;
2. Mapping and alignment of LW-ERIC and external controlled vocabularies;
3. Editorial organization.

The discussion started with the evaluation if the management process of controlled vocabularies in LW-ERIC should be centralized or distributed.

One of the possibility could be to have a centralized common platform providing different tools such as those for editing or mapping controlled vocabularies. At now there are no tools for the management of thesauri allowing the "centralized solution", but perhaps open tools such as TemaTres can be customized in order to add a layer for communication. In any case (centralize or stand-alone tool) we need a pipeline/workflow prior to deployment that ckecks. If the tools are centralize, the pipeline can be integrated in the edition process. If stand-alone the pipeline needs to be integrated during the submission process.

Participants decided to produce a table with the specific requirements that a knowledge organization tool must have in order to satisfy the LW-ERIC needs.

The main requirements for the different phases of vocabulary implementation, curation and publication are as follow:

1. Vocabulary implementation and curation
   - Supporting international standards and formats (SKOS/RDF)
   - Batch import support
   - Collaborative workflow
   - User account control (AAI support)
   - Multiprojects
   - Multilingualism
   - Versioning
   - Triggering capabilities
   - Provenance
   - Mapping and alignment functionality

2. Vocabulary publication

    - Web Interface
    - Search capabilities
    - Export
    - Sparql endpoint
    - API

Moreover, it was also considered the requirement in terms of costs (Open source or proprietary tools).

The aim of table was to compare existing tools (not only the tools at now used in the LW-ERIC community but also at international level) and to assess which one could be used and further customised for the LW-ERIC purpose.

There are a variety of options for publishing controlled vocabularies, the tools we have considered are:

- TemaTres
- Themas Back Bone Thesaurus (BBT)
- Semantic MediaWiki
- PoolParty
- TopBraid EDG—Vocabulary Management
- CESAB ThesauForm
- VocBench

For each tool, the working group reported information related to the specific requirements listed above, with findings/results summarized in spreadsheet form.

The working group decided to organize a meeting (end of january or february) for performing tests on the proposed tools. The meeting will involve editors and ICT experts in order to evaluate both the performance aspects and also the usability for the scientific community involved. After this comparison, a solution for the implementation, curation and alignment of standardized terminological resources will be proposed at the Executive Board. Then the decision of the Executive Board, a plan of activities including a series of working meetings will be established for the next year in order to achieve the planned goals.

Ontologies

Introduction

To address the today's ecological challenges, it is necessary to use data coming from different disciplines and providers. Thus, discovery and integration of data, especially from the ecological domain, is highly labour-intensive and often ambiguous in semantic terms. To improve the location, interpretation and integration of data based on its inherent meaning, ontologies can help in harmonizing and enriching descriptions of data providing a formal mechanism for the definition of terms and their relationships.

Ontologies are representations of the knowledge within a domain of interest, defined via the terminology (concepts) used within the domain and the properties and relationships among domain objects (Baader et al., 2003). In this way, ontologies represent one enabling mechanism for providing more comprehensive data discovery, integration (Jones et al., 2006) and analysis.

In the last years, in the framework of LifeWatch, research groups and projects focusing in the monitoring and analysis of ecosystem properties have increasingly put effort into the development of semantic resources mainly based on core ontologies.

LW-ERIC aims to define a common strategy to develop technology to discover, access, integrate, and analyze distributed ecological information.

Tour de Table Presentations

The session began with a tour de table with short presentations by national nodes working and/or using ontologies. The aim was to provide an overview of the current practices and tools within each node and to get the background for following discussions.

Paul Martin, on behalf of LifeWatch Netherlands, introduces the work done in the framework of the ENVRIplus project: Open Information Linking for Environmental science research infrastructures (OIL-E). OIL-E is intended to provide a framework for semantic linking between different RI standards and vocabularies.

Using the archetypes of the ENVRI Reference Model to produce an upper ontology for RI specifications, OIL-E provides a linking model for describing the overlaps between the different metadata schemes used by RIs to describe their resources, as well as the semantic mappings used to convert between schemes.

Using OIL-E, the ENVRIplus project has built a knowledge base describing the semantic landscape of environmental science RIs in Europe, capturing information about metadata schemes, ontologies, thesauri and other controlled vocabularies used by RIs and helping to navigate the semantic bottlenecks facing the establishment of an open science commons in Europe and beyond.

Xiaofeng Liao introduced the approach proposed for the Semantic annotation of Documents developed in ENVRIplus.

To increase the degree of automation and reduce the human expertise involved in semantic annotation, different fully-automatic or unsupervised approaches are investigated to find their advantages and disadvantages. Briefly, most of these approaches accomplish full automation via utilizing similarity in the text from different aspects or at different levels, either structure similarity, linguistic similarity, or semantic similarity. Some of the most notable approaches are, bootstrapping, clustering, wrapper induction, graph ranking, etc.

LifeWatch Italy introduces its own Core Model, based on a customization of the OBOE Core, for the semantic description/capture of basic concepts and relationships in ecological studies. This framework ontology is based on 7 main concepts (classes) as Domain, Entity, Observation, Characteristic, Measurement, Protocol, Standard, providing a structured yet generic approach for semantic data annotation, and for developing domain-specific ecological ontologies as the Phytoplankton Trait Ontology (PhyTO). To date, LifeWatch e-Infrastructure stores and manages data and metadata using an mix of Database Management Systems (the Relational MySQL and the NoSQL MONGO DB); for the purpose of the study case, we selected the VIRTUOSO Triple Store as semantic repository and we developed different modules to automate the management workflow.

A first software module has been developed to allow the data annotation with classes, subclasses and properties of the PhyTO (i.e. Semantic Annotation). The designed module allows to map metadata and data stored in the LifeWatch Data Portal with the OWL schema of the PhyTO and to produce .rdf output files. A second developed module uses as input the .rdf files and store the data in the VIRTUOSO Graph to make them available for the semantic search. Moreover, a user-friendly search interface (i.e. Java Portlet) has been implemented to retrieve annotated data with queries suggested by the data users.

This approach facilitates data discovery and integration, and can provide guidance for, and automate, data aggregation and summary.

LifeWatch Greece introduces the evolution of the novel mapping and transformation tools that have been implemented in FORTH and used in the LW Greece infrastructure. Moreover they proposed the adoption of a data aggregation and provision workflow and the corresponding tools that implement it. Finally they presented CRMSci a CIDOC-CRM (http://cidoc-crm.org/) based scientific observation ontology that can be used as the basis for the creation of the semantic model of LW ERIC.

Discussion

After the tour de table the main point of discussion was the creation of a new LifeWatch ERIC model vs using an existing one and the future step to do.

Considering Semantic a crucial topic for the LifeWatch ERIC infrastructure is important to start to talk about a LifeWatch ERIC model. The integration between OIL-E (LW Netherlands / ENVRI), CIDOC-CRM (LW Greece) and LW Italy core/domain ontologies could be a good first approach towards a generic LifeWatch Ontology that could semantically integrate all projects under the LifeWatch umbrella.

Another interesting approach towards integrating the different datasets collected in LifeWatch national nodes would be to annotate them using schema.org and other standard vocabularies (such as those we presented during the workshop, e.g. VANN, VoID, DCAT…), to provide a central dataset search tool. This can be deployed either as annotations on each LifeWatch node website, or in a central repository in lifewatch.eu main site. Again, the approach could start by using existing vocabularies to characterize datasets, but with the aim to further enrich them using a LifeWatch Ontology that would give added value to users of the RI.

The participants agree to introduce these approaches to the Executive Board and to prepare a plan of activities for testing and verifying the efficiency of the proposed solutions. A series of working meetings (face2face and virtual) will be planned for the next year to continue the work.