Report

Ontology & Semantic Web for Research

Lecce, 11-14 July 2017

Index

Introduction 4

Participants 5

Session I – Usages of semantic resources in Research Infrastructures 6

Pre-workshop task summary 6

Presentations and discussion notes. 6

Semantics for the Analysis and Experimentation on (continental) Ecosystems: AnaEE (Christian Pichot, INRA) 6

Terminology Supported Data Archiving and Publication in PANGAEA (Markus Stocker, University of Bremen) 7

Semantic Resources in LifeWatch Italy (Ilaria Rosati & Caterina Bergami, LifeWatch Italy) 8

Semantic Data Management in the AquaDiva Project (Friederike Klan, Friedrich-Schiller-University Jena) 8

Semantic monitoring data process description in LTER within SERONTO and EnvThes (Barbara Magagna, LTER) 9

The Vi-SEEM e-Infrastructure project (Valentina Vassallo, The Cyprus Institute) 9

Session II – Alignment of vocabularies and ontologies 11

Pre-workshop task summary 11

Presentations and discussion notes. 12

Alignment of the AnaEE thesaurus, and ontology. C. Pichot, INRA 12

Bridging multiple domains through an environment ontology: the value of continuous semantic interoperation and collaborative development. P. Buttigieg, Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung 13

The LifeWatch Italy semantic approach. N. Fiore, LifeWatch Italy 13

Observable characteristics - existing approaches and arising problems in EnvThes and other related vocabularies. B.Magagna, LTER 14

Future Work 14

Session III - Semantic interoperability and discoverability 15

Session IV – Services for semantics 17

Pre-workshop task summary 17

Presentations 17

Selecting and Customizing Ontologies with JOYCE (F. Klan, Friedrich-Schiller-University Jena) 17

Semantic enablement of geospatial metadata: Going full circle (C. Fugazza, IREA - CNR) 18

WordNets for Modelling Word Meaning in Linguistics and Cognitive Psychology Research and The The CLARIN Concept Registry (E. Hinrichs, CLARIN Research Infrastructure) 19

Open information linking across environmental research infrastructures (P. Martin, University of Amsterdam; ENVRIplus project) 19

IndexMed consortium for data mining in ecology: How to build graphs and mine heterogeneous data for environmental research? (Romain David, IMBE - CNRS) 20

Introduction

Biodiversity, as many other scientific fields, relies on the integration of data coming from multiple sources and spanning multiple scales (spatial and temporal). As the volume of available data is increasing, several projects and infrastructures, are developing services for better managing this information, making it accessible, available and re-usable in order to create new knowledge. Alongside, community standards, tools, services and governance models were developed to facilitate data and system interoperability. A common strategy is to exploit semantic resources, such as metadata, vocabularies and ontologies, to support interoperability among emerging data infrastructures. The workshop, resulting from the collaboration between LifeWatch and EUDAT, has been a chance to discuss common approaches, used tools and existing solutions, not only for the biodiversity communities, but also for as many scientific domains as possible. The goal was to continue the discussion started last year during the LifeWatch workshop "Thesauri & Semantics in the Ecological Domain" (http://www.servicecentrelifewatch.eu/documents/28189/689756/ThesauriSemanticsWorkshopReport.pdf/3e4af94c-04be-40fc-a018-9ac1d556384b) and during the EUDAT workshop in Barcelona (https://www.eudat.eu/events/trainings/co-located-eudat-semantic-working-group-workshop-9th-rda-plenary-barcelona-3-4) to propose good practices for semantic resource development, interoperability and discoverability, the definition of authoritative tools and facilities for the scientific community.

Participants

Name	Surname	e-mail	University, Institution, Research Centre, etc
Alberto	Basset	alberto.basset@unisalento.it	DiSTeBA, University of Salento
Alessandro	Oggioni	oggioni.a@irea.cnr.it	LifeWatch Italy, CNR - Institute for Electromagnetic Sensing of the Environment (IREA)
Angela	Boggero	a.boggero@ise.cnr.it	LifeWatch Italy, CNR-Institute of Ecosystem Study
Bachir	Balech	balechbachir@gmail.com	LifeWatch Italy, IBIOM-CNR
Barbara	Magagna	barbara.magagna@umweltbundesamt.at	EAA
Cataldo	Pierri	cataldo.pierri@uniba.it	LifeWatch Italy, CNR-IBAF
Caterina	Bergami	caterina.bergami@ismar.cnr.it	LifeWatch Italy, CNR-IBAF
Christian	Pichot	christian.pichot@inra.fr	INRA
Cristiano	Fugazza	fugazza.c@irea.cnr.it	IREA-CNR
Doron	Goldfarb	doron.goldfarb@umweltbundesamt.at	Environment Agency Austria
Elena	Stanca	elena.stanca@unisalento.it	DiSTeBA, University of Salento
Erhard	Hinrichs	Erhard.hinrichs@uni-tuebingen.de	TŸbingen University; CLARIN
Friederike	Klan	friederike.klan@uni-jena.de	Friedrich-Schiller-University Jena, Germany
ilaria	Rosati	ilaria.rosati@unisalento.it	LifeWatch Italy, CNR-IBAF
Marie	Hinrichs	Marie.hinrichs@uni-tuebingen.de	TŸbingen University; CLARIN
Markus	Stocker	mstocker@marum.de	University of Bremen
Naouel	Karam	naouel.karam@fu-berlin.de	Freie Universit_t Berlin
Nicola	Fiore	nicola.fiore@unisalento.it	LifeWatch Italy
Oya	Beyan	beyan@dbis.rwth-aachen.de	RWTH Aachen
Paolo	Tagliolato	paolo.tagliolato@gmail.com	LifeWatch Italy, CNR - Institute for Electromagnetic Sensing of the Environment (IREA)
Paolo	Colangelo	paolo.colangelo@uniroma1.it	LifeWatch Italy, CNR-ISE
Paul	Martin	p.w.martin@uva.nl	University of Amsterdam
Philip	Trembath	phtr@ceh,.ac.uk	Centre for Ecology & Hydrology, UK
Pier Luigi	Buttigieg	pbuttigi@mpi-bremen.de	Alfred-Wegener-Institut, Helmholtz-Zentrum fŸr Polar- und Meeresforschung
Romain	DAVID	romain.david@imbe.fr	IMBE - CNRS
Stefano	De Felici	stefano.de.felici@uniroma2.it	LifeWatch Italy, CNR-IBAF
Tommaso	Di Noia	tommaso.dinoia@poliba.it	Polytechnic University of Bari
Valentina	Vassallo	v.vassallo@cyi.ac.cy	The Cyprus Institute; VI-SEEM Project
Yann	Le Franc	ylefranc@esciencefactory.com	e-Science Data Factory
Zhiming	Zhao	z.zhao@uva.nl	University of Amsterdam

Session I – Usages of semantic resources in Research Infrastructures

Editor: Paolo Tagliolato, CNR ISMAR - LifeWatch Italy ICT WG

Pre-workshop task summary

- Which semantic resources did you use or are you using and in which scientific domain?
  - Vocabularies, thesauri and ontologies
  - Biology, medicine, marine ecology, functional ecology, social science, environmental science, research infrastructure analysis, cultural heritage
- How and where (e.g. research infrastructure and project, data platform, etc.) did/do you use them?
  - Document annotation/indexing
  - Metadata creation/annotation.
  - Data curation and sharing
  - Data mining, Information Retrieval
- Which are the advantages and disadvantages in using semantics in your research?
  - Pros: Common language; Disambiguation; Interoperability; Enablement of discovery and reusability; mobilising datasets,
  - Cons: effort required to develop such resources and to get users to use them; multilingualism is necessary but requires additional effort; unclear success stories; automatic annotation far from perfect, additional effort is required to check it; diversity of formalism hampers usage;
- Which (national or international?) metadata standards do you follow?
  - DCAT; DCAT_AP_IT; ISO 19115; ISO 19139; CF convention; TDWG recommendations; OBO principles; Genomic Standards Consortium standards; OGC standards; RDA; MARC; JATS; CMDI; INSPIRE-EMF; INSPIRE ISO 19139; OAI-PMH; SSN; SensorML; Dublin Core; DIF; Darwin Core; LifeWatch Italy dataportal metadata profile; RNDT; Eagle (epigraphic resources), STARC (digital provenance of 3D cultural resources)
- Which metadata language, ontologies and controlled vocabularies do you use
  - DCAT, SSN, OBOE,OWL-Time, ORG, FOAF, GCMD, ChEBI, GACS, re3data, Nature, QUDT, EnvThes, NVS, ANAEETHES, LifeWatch Thesauri, Geonames, GEMET, AGROVOC, …
- How is data harmonized and standardized, and which (national or international?) protocols and standards do you follow?
  - Community and consortia recommendations and conventions; iterative process; following INSPIRE directive;
- At what stage of the data lifecycle is the semantic aspect introduced?
  - (All stages)

Presentations and discussion notes.

Semantics for the Analysis and Experimentation on (continental) Ecosystems: AnaEE (Christian Pichot, INRA)

AnaEE (Analysis and Experimentations on Ecosystems) is a R.I. currently in development, with some already deployed components, with the example of AnaEE-France (https://www.anaee-france.fr/en/), reported in this presentation. The infrastructure is distributed and comprising heterogeneous resources. To achieve interoperability, semantics is considered for annotation of databases and modelling platforms ("tools to analyse, synthesise and project"). Exploited semantic resources are thesauri (**AnaEE-France** specific and existing Thesauri, in particular **GEMET**[1] and **AGROVOC**[2] from Agroportal and **EnvThes**[3] from LTER) and ontologies (in particular an ontology based on **OBOE**[i] [4] to **annotate observations**, and **SSN**[ii] [5] to **annotate sensors**). Work in progress is the access portal to AnaEE resources through ***RDF versions of metadata*** currently available in standard (XML) formats like ***ISO-19139***[iii] ***and EML***[6].

Open questions from discussion: alternatives to OBOE for modelling observations towards interoperability **OGC SWE**[iv]**/O&M**[v] or **SSN** for modelling observations emerges (M. Stoker)

Terminology Supported Data Archiving and Publication in PANGAEA (Markus Stocker, University of Bremen)

PANGAEA (https://www.pangaea.de/about/) is a Multidisciplinary data publisher for environmental sciences, whose information system hosts 360 K citable data sets (DOI) and 11 B data items within **relational database** and data **warehouse**, offering **search engine (ElasticSearch**[7]), an **editorial system** and **APIs**.

The terminology catalogue (TC) here discussed is a component for dealing with the over 140K measurements and observation types (parameters) and complex compound concepts in use within hosted data. TC is exploited in **ingestion, archiving, access and dissemination activities**. It is synchronized with external terminologies as well. TC is based on a relational model, in consideration of the fact that RDF support would be too tedious, but it provides an **API to serve contents in RDF**.

Rules for concatenating terms of different categories (e.g. Quantity kind, features, quantity unit) are proposed in order to deal with compound concepts*.

Takeaways: Approach to terminology and semantics based on relational model; Better consistency in archived data and **improve search**; Improve term synchronization workflows between systems.

The TC is used for **metadata enrichment**, **access to data** (e.g. including ***synonyms***, ***expanding to broader terms***). Mapping of TC terms to external vocabularies is a strategy to facilitate dissemination.

Discussion Notes. Is TC is ***available for the public? Currently not***, but it could be considered to open it through some API if a broader community would be interested.

[1] https://www.eionet.europa.eu/gemet/en/themes/

[2] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[3] http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn

[4] https://semtools.ecoinformatics.org/repository/dev/oboe

[5] https://www.w3.org/2005/Incubator/ssn/ssnx/ssn

[6] https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html

[7] https://www.elastic.co/products/elasticsearch, source repository at https://github.com/elastic/elasticsearch (Apache Licence 2.0)

Semantic Resources in LifeWatch Italy (Ilaria Rosati & Caterina Bergami, LifeWatch Italy)

LifeWatch Italy is focussing on the scientific domain of biodiversity and ecosystems. Given the lack of semantic resources for functional ecology, LW-ITA fostered the development of **skos thesauri** on **functional traits** of several groups of **aquatic organisms** such as phytoplankton, zooplankton, fish, macrozoobenthos, macroalgae but also **thesauri on alien species**, **endemism**, **genomic** and **barcoding**.

Thesauri have been developed and managed, through a collaborative process, by working groups of: editors (domain experts with the responsibility of contents); ICT experts (supervising technological aspects, semantic modelling and providing software tools for supporting thesauri composition and publication); validators (domain experts reviewing contents).

The work is supported by the **TemaTres editor** (open source software), exposing also resources through an integrated **SPARQL endpoint**, as well as **LD** and through **API** (useful for integration with e.g. thesaurus publishing interface).

Thesauri are currently used in the LifeWatch Italy Data Portal (http://www.servicecentrelifewatch.eu/catalogue-of-resources) for the **definitions of Metadata and Data Schema elements**, to fill **metadata elements values** and as building blocks for the on-going construction of the **LifeWatch ontology** (based on **OBOE**). Ingestion UI provides a way for **harmonizing tabular data** uploaded by users, through linking original **column names with thesauri Concepts**, an information that is returned in publishing phase through the **dataportal APIs** (json representation).

Next steps for LW-ITA thesauri will be their extension to other functional traits and further development of multilingualism, and mapping with other thesauri (EnvThes, NVS, TopThesaurus).

Discussion Notes. LW-ITA thesauri are not currently exploited externally from the LW infrastructure, but they are publicly accessible and referenceable. Work in progress, as said, is the reciprocal mapping with other thesauri through established collaborations among editor teams.

Semantic Data Management in the AquaDiva Project (Friederike Klan, Friedrich-Schiller-University Jena)

AquaDiva (http://www.aquadiva.uni-jena.de/) is a project focusing on roles of water and biodiversity. Infrastructure is based on Bexis (BExIS 2 – a Flexible Data Management Platform for Biodiversity[8]), whose features are data upload, metadata management, rights management, data publishing to GFBio[9], data search.

In the **data upload** phase the system enables the user to **annotate tabular data with units and data types**. At Search level, semantic search is enabled by **mapping RDBMS data to the RI knowledge base**, modelling data according to **OBOE**. Characteristics of OBOE observations range in several related ontologies: **oboe-chemistry, oboe-temporal, ENVO, NCIT-module, OBI, ChEBI-light-module**. Reasoning and SPARQL endpoint are enabled by the **QUEST reasoner software**[vi] by means of which data are made available in a virtual Abox of the knowledge base. The system makes possible to query data by transforming sparql queries into sql queries. The AquaDiva project developed, on basis of

[8] http://bexis2.uni-jena.de/

[9] https://www.gfbio.org/

research questions elicited by the users community, a **user interface for facilitated access to data**: the UI is able to **transform keywords specified by user into SPARQL queries** with a predefined structure.

Discussion notes. UI is open software.

Additional notes. **QUEST** is part of the **Ontop framework** (http://ontop.inf.unibz.it/) for the enablement of the Ontology Based Data Access (OBDA) scenario. The software is open source and is available at the github repository https://github.com/ontop/ontop.

Semantic monitoring data process description in LTER within SERONTO and EnvThes (Barbara Magagna, LTER)

The Long Term Ecosystem Research (LTER) (www.lter-europe.net) is a network comprising in Europe more than 500 sites and 24 national networks, with 100 institutions and more than 1000 scientists involved. The Socio-Ecological and Ecological Research and Monitoring ONTOlogy (**SERONTO**[vii]) was conceived during the FP6 project AlterNet (2004-2006) as a framework for management and integration of in-situ monitoring data (observations and measurements) within the network. The core of Seronto was developed with different aims, among which adhering to w3c standards and being compatible with the Ecological Metadata Language (EML). The ontology was developed by the community with an iterative process involving a working group proposing solutions, and then involving the entire community in the discussion and decision phase, being consensus of first importance for the initiative. This governance model revealed itself as very time consuming, and the initiative ran out of time.

The common basis developed during this initiative was instead exploited for the construction of **EnvThes** thesaurus, which is now available and in use by several projects for data **harmonization** (LTER), **data exchange** (eLTER) and **service validation** (ENVRI Plus).

LTER DEIMS system (http://data.lter-europe.net/deims/) enables users to **annotate metadata** with EnvThes Concepts for keywords during the **curation phase** (**metadata creation and ingestion**).

EnvThes is maintained by an Editor Team and technologically supported by the use of the **TopBraid skos thesauri editor software** hosted at CEH (http://evn.ceh.ac.uk/). The same platform offers a SPARQL endpoint (http://vocabs.ceh.ac.uk/evn/tbl/sparql) and Linked data access to EnvThes.

Notes: TopBraid Enterprise Vocabulary Net (EVN) is a proprietary web based platform for managing thesauri, developed by TopQuadrant.

The Vi-SEEM e-Infrastructure project (Valentina Vassallo, The Cyprus Institute)

Virtual Research Environment for regional interdisciplinary communities (Climatology, Life Sciences, and Digital Cultural Heritage) in Southeast Europe and the Eastern Mediterranean (Vi-SEEM) is a 36 months project started at the end of 2015. The scope of ViSEEM is to merge Southeast Europe and the Eastern Mediterranean regions under a service oriented e-Infrastructure built during the last decade thanks to the efforts of various projects carried out within the two areas. Current most important services are ChemBioServer (filtering, clustering and visualization of chemical compounds), Live Access Server (access, visualization and post-processing of geo-referenced scientific and climate data), Clowder (Digital Culture Heritage repository with interactive visualizations).

For data management and access, a work in progress is the development of a cross-disciplinary semantic solution for different resources. A case study regards **description**, **aggregation** and **retrieval of museum datasets** from different subjects (Banja Luka, Republika of Srpska, Bosnia and Herzegovina). **Metadata** for the different contents are created following **STARC**[viii] **schema**, which is able to integrate information regarding the real objects and their digital “surrogates” (3D models, photographic documentation, digital texts, etc.) and also digital resource provenance.

Discussion notes: linking provenance information and VRE workflows is a work in progress.

Session II – Alignment of vocabularies and ontologies

Editors: Nicola Fiore, UniSalento - LifeWatch Italy, Barbara Magagna, Environmental Agency Austria

Ontology/Vocabularies alignment is the process of determining the commonality between classes and concepts from different Ontologies/Vocabularies.

(Adam M.; Vodden, Peter N. 2016)

Pre-workshop task summary

How to reuse, adapt and extend semantic models and instances from existing resources to build your own ontology?

- To extend an existing model it is important to understand what is your aim for building a new ontology and if the terms you wish to add are not existing in previous ontology

How to be supported in this process?

➢ **Ontology Marketplace**: supplier/user

Def. *The "marketplace" can be seen as a complex of conceptual, organizational and technical measures that ensure an effective and efficient information exchange between suppliers of semantic assets (ontologies) and their consumers – owners of the practical Use Cases.*

How to reuse, adapt and extend semantic models and instances from existing resources to build your own ontology?

- The goal when building a new ontology should be to avoid "an ontology of everything" but to build a minimal core ontology that can form the basis of many solutions. In the design analysis it is important to **explore and compare the state of the art of the existing ontologies and thesauri** in order to identify the commonality between concepts.
- Ontology lookup service or repositories with all vocabularies and ontologies of the relevant domains included in it could help to interlink more efficiently your specific vocabulary with other semantic resources to be able to bridge between a more generic approach to specific ones according to the needs of the scientists.

Do you use semantic resources aligned with an upper ontology?

If yes, which one?

Def. An **Upper Ontology** (also known as a **top-level ontology** or **foundation ontology**) is an ontology which consists of very general terms (such as "object", "property", "relation") that are common across all domains. An important function of an upper ontology is to support broad semantic interoperability among a large number of domain-specific ontologies by providing a common starting point for the formulation of definitions. Terms in the domain ontology are ranked "under" the terms in the upper ontology, and the former stand to the latter in subclass relations.

DOLCE and OBOE are the most upper Ontologies used.

Which kind of services or tools could be useful to avoid concept duplications?

- **Ontology Lookup Services**
- **Upper Ontologies**
- **Ontology Design Patterns** for specific use cases

Are you familiar with the existing ontology design good practices for reuse in other scientific domains?

- Missing overview of existing ontologies design practices -> RDA VSIG (VSSIG) knowledge to be included in marketplace?
- Create together list of wide used basic ontologies and models: O&M, O&M light, OBOE, (DOLCE), ENVO, SSN, SensorML, Observable properties, Complex properties, … to be presented shortly Friday morning (working sessions)

Are you aware of the existing Biomedical initiatives and approaches for ontology interoperability such as the OBO foundry initiative and the related MIREOT approach for importing and reusing existing ontological concepts (Courtot and al., 2011)?

- There is not a general overview on this initiatives.
- EUDAT semantic working group – intends to offer a training on this? The scientist is lost without good overview and training.

Presentations and discussion notes.

Alignment of the AnaEE thesaurus, and ontology. C. Pichot, INRA

Exploited semantic resources from AnaEE are thesauri (AnaEE-France specific and existing Thesauri, in particular **GEMET** and **AGROVOC** from Agroportal and EnvThes from LTER) and ontologies (in particular an ontology based on **OBOE** to annotate observations, and SSN to annotate sensors).

AnaEE has aligned the existing **AnaEE Ontology** (based on OBOE) with the AnaEE Thesaurus, and the AnaEE thesaurus with the GEMET and AGROVOC Thesauri. The Ontology alignment has been developed using the skos:exactMatch properties when extending the ontology within webprotege. The vocabularies alignment has been developed using AML (AgreementMakerLight) ontology mathing systems.

The AnaEE thesaurus is yet in a construction phase due to limited resources, actually it is composed by 3.320 concepts: 400 aligned with Agrovoc and 280 with GEMET.

Discussion notes. Vocabularies to align with methodology (alignment criteria)

Bridging multiple domains through an environment ontology: the value of continuous semantic interoperation and collaborative development. P. Buttigieg, Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung

The Environment Ontology (ENVO; http://www.environmentontology.org) is a resource and research target for the semantically controlled description of environmental entities.

ENVO is developed in the Web Ontology Language (OWL) and employ templating methods to accelerate class creation. Steps are taken to better align ENVO with the Open Biological and Biomedical Ontologies (OBO) Foundry principles and interoperate with existing OBO ontologies. Further, text-mining approaches has been applied to extract habitat information from the Encyclopedia of Life and automatically create experimental habitat classes within ENVO. ENVO offers representations of habitats, environmental processes, anthropogenic environments, and entities relevant to environmental health initiatives and the global Sustainable Development Agenda for 2030. Several branches of ENVO have been used to incubate and seed new ontologies in previously unrepresented domains, such as food and agronomy. ENVO has been shaped into an ontology which bridges multiple domains including biomedicine, natural and anthropogenic ecology, 'omics, and socioeconomic development.

Discussion notes. Github like discussion and interoperation promoting alignment. For modelling of LTER observations, including methods and provenance information, a cooperation with ENVO is encouraged.

The LifeWatch Italy semantic approach. N. Fiore, LifeWatch Italy

LifeWatch, the European e-Science Infrastructure for Biodiversity and Ecosystem Research, is an ERIC since March 2017. In the last years, the Italian node has developed a complete infrastructure to support the entire data Lifecycle, from the collection to the analysis of the data in specific Virtual Research Environment. The team has designed and is experimenting an architecture based on a set of domain ontologies aligned with the **OBOE core ontology** and a set of thesauri (LifeWatch Italy thesauri - **thesauri on alien species**, **endemism**, **genomic** and **barcoding**). All the datasets shared by the community are annotated using this approach and made available in a Virtuoso triple store that allows the rdf export and the analysis thought a SPARQL end-point. The team is working now on a specific study case in the phytoplankton domain to design and build user-friendly interface for data analysis.

Discussion notes.

Which are the countries involved in LifeWatch? LifeWatch-ERIC has been funded by 8 Full Members (Belgium, Greece, Italy, Portugal, Romania, Spain, Slovenia, The Netherlands), also hosting the related national nodes, and constituting the Distributed Centres of the infrastructure.

Could we think to use the phytoplankton study case to compare the different semantic approaches? Yes.

Observable characteristics - existing approaches and arising problems in EnvThes and other related vocabularies. B.Magagna, LTER

EnvThes (http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn) compiles a set of terms relevant to describe and annotate data resulting from long (and short) term ecosystem research. It is used to select keywords for the annotation of datasets, sites and data products in DEIMS: https://data.lter-europe.net/deims/ and to select proper common terms for translating local parameter naming (e.g. in a data table or SOS service) to common parameter names agreed in the (LTER) community. The EnvThes *measure* concept is a compound term as used from the LTER researcher (e.g. concentration of sulphate in soilwater), but additionally atomic concepts like property (e.g. concentration), object of interest (sulphate), matrix (soilwater), tool (lysimeter) are used to provide powerful searching and easy discovery. It is planned to develop extended parameter discovery ontology where the complex and the atomic concepts are linked to each other. A second ontology will be built upon it, based on SERONTO or other suitable observation ontologies, to include detailed documentation of the sampling and observation methods applied in the measurement.

Discussion notes.

How to atomize complex properties – this seems not to be always clear. In AquaDiva project have similar problems were encountered. It could be problematic using EnvThes terms as instances in the ontology, as a Thesaurus always deals with concepts and not real things in the world. But this depends on the type of classes to be instantiated. A repository, like the envisaged EcoPortal which gathers all relevant thesauri and ontologies, but also SKOS reference lists in the ecology domain, would support research for LTER to a great extent.

Future Work

At the end of the session all the participants have agreed to work on the specific Phytoplankton study case (file attached) described by LifeWatch Italy. Each Research Infrastructure/ Research Group will model the Study Case with the developed Ontology and an alignment between the different concepts will be discussed as final result of the session. The goal will be to write a paper on this alignment. The first diagrams produced during the working group session are attached.

Session III - Semantic interoperability and discoverability

Editor: Yann Le Franc, PhD, EUDAT

This session aimed at discussing about two major issues impacting the development of semantic tools and services: the **discoverability of vocabularies and vocabulary services** and the **interoperability of vocabulary services APIs**. The first part of the session was focused on an introductory presentation of EUDAT infrastructure, services and the current development of semantic services within EUDAT with the semantic annotation service B2NOTE [10]. The presentation then focused on the two issues we are tackling to build a multi-disciplinary annotation service and our approach within EUDAT to address these issues.

The first issue we are tackling is related to the discoverability of the existing semantic resources within a domain and across multiple domains. The short survey filled in by the workshop participants prior to the event, showed that the main ways to discover ontologies within a domain is through networking within the community, publications or using google. For the semantic resources outside of the domains, only few are using ontologies from different domains. However, the answers show a trend toward multi-disciplinary integration especially in the context of Biodiversity. Indeed, this domain needs information coming from multiple sources and multiples domains (chemistry, biomedical, biology, earth science…). Many semantic resources exist and are scattered through the web, which makes it difficult to have an overview the current wealth of resources. To address this issue, we are building within EUDAT a proof-of-concept implementation for a semantic resource look up service. This service should allow to aggregate semantic resources from all scientific domains within a catalogue, and generate a multi-disciplinary semantic index of concepts/terms. This index can thus be used by semantic services, scientists, knowledge engineers, ontologists and become a valuable resource for analysing the content of these repositories.

The second issue we are encountering is related to API interoperability of the semantic resources, i.e. vocabularies/ontologies and vocabulary/ontology services. Indeed, the large diversity of semantic repositories comes at the price of a large diversity of APIs to access their content. This API diversity hampers the scalability of a solution for harvesting the large number of existing semantic repositories. Different existing efforts are proposing technical solutions to enhance web API interoperability, such as the OpenAPI initiative[11], the W3C Hydra community[12] and the smartAPI approach[13]. The short survey showed that the vast majority of the participants were not aware about these initiatives and were eager to investigate them. We then presented the conceptual design of an automated Information Harvester that could be used to harvest the various resources using the current approaches for API interoperability.

[10] https://b2note.bsc.es

[11] https://www.openapis.org/

[12] https://www.w3.org/community/hydra/

[13] http://smart-api.info/website/

The work presented is supported by an international collaboration effort, initiated during a workshop organized by the EUDAT Semantic Working Group in Barcelona (April 3-4, 2017)[14]. This workshop focused on addressing these two issues and three initial working groups were created. These groups focus on the following aspects: defining a common minimal metadata set for semantic resources, metadata and API interoperability between the vocabulary services and design of a semantic marketplace. This effort is now continued within the RDA Vocabulary and Semantic Service Interest Group[15]. The interest group will meet in Montreal in September during the next RDA plenary.

This introduction was followed by three practical examples of application for data interoperability and data discoverability using semantics:

- Markus Stocker presented his work related to the semantization of sensor descriptions for marine biology (ESONET Yellow pages), and how such transformation can be used to harmonize the description of sensors and boost interoperability of the different data repositories using these sensors in the context of the FixO3 observatories.
- Oya Beyan talked about the integration of semantics within data lakes to enhance the discoverability of the data sets, and highlighted the need for vocabulary discoverability and API interoperability.
- Christian Pichot described the approach of semantic interoperability within the ANAEE community and how the semantic annotation of the datasets and databases supported semantic interoperability and discoverability.

These three practical presentations were followed by the presentation of the current implementation of the semantic look up service, by Doron Goldfarb. The presentation provided an overview of the existing vocabulary repositories and of the initial design of the semantic look up service and of the Information Harvester. Then it introduced our initial implementation to address the automation of harvesting the information for the semantic index from REST APIs, using a JSON/JSON-Path approach. This approach allows to create nested queries necessary to extract the needed information. The initial results of harvesting using this approach were then presented. The harvesting was performed from Bioportal, EBI-OLS and Agroportal and gathered more than 13 million terms, including in total almost 9 million terms with unique IDs.

This session was concluded by a discussion regarding our approach, the challenges ahead and the potential extension of the semantic look up services with marketplace functionalities. This discussion continued the following day, during a brainstorming session, to gather additional functionalities that could be built upon the semantic look up service. The results of this brainstorming session have been gathered into a google document[16] and will be used as support for further discussions.

[14] https://eudat.eu/events/trainings/co-located-eudat-semantic-working-group-workshop-9th-rda-plenary-barcelona-3-4

[15] https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html

[16] https://docs.google.com/document/d/1ejkI8RkXCT5I7QH-B-BV39HC7aoJgEAOVuVga8uyi1o/edit?usp=sharing

Session IV – Services for semantics

Editor: Paolo Tagliolato, CNR ISMAR - LifeWatch Italy ICT WG

Pre-workshop task summary

Which kind of tools are deployed or implemented to support a semantic approach?

Tools for semantic annotation of texts; concept registry; RDF portrayal (RDF web viewer); metadata creation tool with embedded semantic annotation mechanism suggesting users terms from controlled vocabularies through runtime SPARQL queries; skos vocabulary editors; UI to SPARQL endpoints with predefined queries to ease user work.

Which kind of services are deployed or implemented to support a semantic approach?

Web services to support semantic annotation and semantic search; SPARQL endpoints and triple stores with spatial search enabled; services for ontology recommendations.

How is the semantic content made available? Through an API, a user interface or a SPARQL endpoint?

All three and Linked Data access

Presentations

Selecting and Customizing Ontologies with JOYCE (F. Klan, Friedrich-Schiller-University Jena)

Jena OntologY Customization Engine[ix] (JOYCE) (http://fusion.cs.uni-jena.de/fusion/publications_overview/jena-ontology-customization-engine-joyce/) is a tool for selecting and customizing ontologies, and enabling their re-use. It is exploited within the Aqua Diva Project (cf. Presentation in Session 1) to assemble project-specific knowledge from ontologies within BioPortal.

In a first configuration section of JOYCE, terms from project data and documents are matched against BioPortal ontologies, then the application, which includes a conceptual filter, identifies relevant classes and minimizes unintended redundancies, i.e. concept duplicates, as well as irrelevant knowledge. The user can configure the maximum number and the type of objects to be combined (either entire ontologies, extraction-based modules or partition-based modules), the size of the random sample taken in each assembly round, and indicate the importance of each of the 3 optimization criteria coverage, overlap and overhead.

The results view section then provides a list of the ontology (module) combinations suggested by JOYCE sorted by their coverage. For each combination, a list of the assembled ontologies (ontology modules), their number as well as coverage, overhead and overlap for each combination are displayed.

The construction of a functionality for Interactive selection of ontologies, as well as that for ontology merging, is on-going.

The GFBio Terminology Service - a unified interface for accessing heterogeneous terminological knowledge (N. Karam, Freie Universität Berlin)

[Speaker could not attend the workshop]

Semantic enablement of geospatial metadata: Going full circle (C. Fugazza, IREA - CNR)

In the last decade, several initiatives addressed interoperability of geospatial resources (INSPIRE, GEOSS, Digital Earth). Syntactic and structural heterogeneities hampering data have been widely addressed (e.g., OGC web services, INSPIRE Data Specifications). Still, locating the geospatial data of interest is a prerequisite to actual usage. Term “discovery” suggests that this activity may constitute a daunting task. The Italian flagship project RITMARE (Ricerca Italiana per il MARE, Italian research for the sea) (http://www.ritmare.it/) developed solutions for interoperable data and metadata. Management of metadata in RITMARE focused on semantic characterization via association of URIs to information items[x]. The more apparent usage of semantics in geospatial discovery is for multilingual retrieval of resources and for query expansion exploiting hierarchical structure of thesauri. It can be applied to the user search pattern, notwithstanding the schema underlying metadata. Specifying URIs in metadata prevents (linear) increase in the number of queries.

EDI[xi] (http://edidemo.get-it.it) is a web-based tool created in the context of RITMARE initially intended to assist metadata creation. It is schema- and data source-agnostic (completely instructable in this respect). Through the usage of EDI, RDF descriptions are obtained as a by-product of metadata editing.

The software had the limitation that pre-existing metadata could not be imported, because it lacks any semantic information (the URIs). Now the Liftboy application is under development and it allows to import pre-existing metadata into the EDI workflow, exploiting parts of the same instructions already composed for instructing EDI.

EDI is an open source software project licensed under GPL3.

WordNets for Modelling Word Meaning in Linguistics and Cognitive Psychology Research and The CLARIN Concept Registry (E. Hinrichs, CLARIN Research Infrastructure)

The Common Language Resources and Technology Infrastructure (CLARIN) (https://www.clarin.eu/) is an ERIC established in 2012. It supports the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. It offers advanced tools (https://www.clarin.eu/content/services) to discover, explore, exploit, annotate, analyse or combine data sets, and in particular its Concept Registry - CCR (https://www.clarin.eu/ccr, https://openskos.meertens.knaw.nl/ccr/browser/) offering a collection of concepts, identifiable by their persistent identifiers, relevant for the domain of language resources, that constitutes the basis of the semantic interoperability layer of CLARIN.

The CLARIN component metadata provides a framework to describe and reuse metadata blueprints, i.e. components that are description building blocks, containing links to the CCR, that can be grouped into a ready made description format. The Component Registry makes these information available for reuse. Single components naturally maps to RDF Classes, while their elements maps to RDF properties.

Further work is currently done on concepts by exploiting WordNet (www.globalwordnet.org), a large lexical database modelling semantic relations like synonyms, hyponym, etc. and covering several languages, especially English (https://wordnet.princeton.edu/), Polish and German (http://www.sfs.uni-tuebingen.de/GermaNet/), that are interlinked by an interlingual index (ILI) to facilitate translation of word senses.

Open information linking across environmental research infrastructures (P. Martin, University of Amsterdam; ENVRIplus project)

ENVRI (http://envri.eu) is a cluster project for environmental science research infrastructures in Europe. It defined a reference model (RM) (http://envri.eu/rm)– specified with UML diagrams - that describes the architecture, operation and information flows of an 'archetypical' RI.

There are a large number of controlled vocabularies and vocabulary management systems (VMSs) being used by RIs, and also plenty of general etadata schemes. Many of these resources can be used to describe objects concerning RIs.

Purposes of Open Information Linking for Environmental RIs (OIL-E)[xii] are

- Bring data describing RIs together, using a standard vocabulary and comparable architectural models.
- Identify/refine a better methodology to guide semantic linking between different standards for (meta)data and services.

From the ENVRI Reference Model, it was defined the 'hub' ontology at the basis of OIL-E, for establishing links with different standards for describing objects, components, processes, etc.

OIL-E is proposed as a semantic model architectural specification.

Now the ENVRI reference model ontology provides the core of the OIL-E framework. Ontology is available at hfp://www.oil-e.net/ontology/ and is split into rm-core, rm-archetypes, rm-correspondences.

The ENVRI plus knowledge base serves at a Fuseki SPARQL endpoint the OIL-E triples, as well as (work in progress) the semantic landscape of RIs based on RM.

The intent is also to apply reasoning on RIs, based on OIL-E ontologies. At the moment ontology complexity allows for limited reasoning, but modularization on OWL.2 sub-profiles is a possible way to perform better.

Several ideas for tools are in design/development exploiting OIL-E.

IndexMed consortium for data mining in ecology: How to build graphs and mine heterogeneous data for environmental research? (Romain David, IMBE - CNRS)

Ecological analyses involve heterogeneous non-linked data at very different scales, formats, and sources. It is a strongly connected system with many interlinkages with numerous, mixed factors of huge local variability. Moreover, communities involved in these studies are multidisciplinary, but only little time is spent for an inter-disciplinary approach. To extract valuable information from this complex system the proposal is to use graphs. They are constructed considering all the possible objects (sites, species, traits, photos, etc.) present in data as nodes, and all values of their factor as links. A prototype is proposed using Neo4J software (https://neo4j.com/) within IndexMed consortium (http://www.indexmed.eu/). It lets users choose external databases, select nodes and links, then construct the corresponding graph, which can be refined to obtain a graphic representation, and finally export the result that is associated to it a URL. A web service is available for data computing. Examples of data visualization are presented for the use cases of underwater photos and archaeological sites.

Next challenges are: to use multi layer graphs (for different domain layers like traits, species, contexts); to exploit pattern recognition and clustering based on number and strength of links between nodes; to adopt FAIR principle. Current status: first prototype for visualization; community of about 250 involved researchers, mostly in environmental sciences; indexing data web service; years of presentation of the project in each disciplinary communities; workshops done for presentation and conception of graphs.

IndexMed development was helped by ECOSCOPE (http://ecoscope.fondationbiodiversite.fr/fr/), biodiversity data hub, a national RI recognized by the ministry of research, with the aim of facilitating access to observation data and of encouraging complementarity of observations (and interoperability of datasets) through all levels of data life cycle.

Finally, the SemanDiv GDR is introduced, a French research network funded for 2017-2020, devoted to the semantic of biodiversity with the primary objective of contributing to solve semantic heterogeneity of biological and ecological facets of biodiversity. It works on 4 axes: delimitation of the scientific field; development of semantic standards (thesaurus and ontology for biodiversity); inventory of terminological standards; visualization, queries and mapping with databases.

[i] Madin, J. et al., 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics, 2(3), pp.279–296.

[ii] M. Compton, P. Barnaghi, L. Bermudez, R. Garc ́ıa-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. Le Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor. The SSN ontology of the W3C semantic sensor

network incubator group. Web Semantics: Science, Services and Agents on the World Wide Web, 17:25–32, Dec. 2012.

[iii] ISO/TC-211. Geographic information – metadata – xml schema implementation. Standard ISO 19139:2007, International Organization for Standardization, Geneva, 2007.

[iv] OGC. OGC® SWE Common Data Model Encoding Standard. Standard OGC 08-094r1, Open Geospatial Consortium, 2011.

[v] OGC. Geographic information — observations and measurements. OGC Standard: Abstract Specification OGC 10-004r3, Open Geospatial Consortium, 2013.

[vi] Rodrıguez-Muro, M., & Calvanese, D. (2012). Quest, an OWL 2 QL reasoner for ontology-based data access. OWLED 2012.

[vii] van der Werf, D. C., Adamescu, M., Ayromlou, M., Bertrand, N., Borovec, J., Boussard, H., ... & Hammen, V. (2008). SERONTO: a Socio-Ecological Research and Observation oNTOlogy. Proceedings of TDWG, 17-25.

[viii] Ronzino, P., Hermon, S., & Niccolucci, F. (2012). A metadata schema for cultural heritage documentation. V., CApellini (ed.), Electronic Imaging & the Visual Arts: EVA, 36-41.

[ix] E. Faessler, F. Klan, A. Algergawy, B. K¨onig-Ries, and U. Hahn. Selecting and Tailoring Ontologies with JOYCE, pages 114–118. Springer International Publishing, Cham, 2017.

[x] C. Fugazza, M. Pepe, A. Oggioni, P. Tagliolato, F. Pavesi, and P. Carrara. Describing geospatial assets in the web of data: A metadata management scenario. ISPRS International Journal of Geo-Information, 5(12):229, 2016.

[xi] F. Pavesi, A. Basoni, C. Fugazza, S. Menegon, A. Oggioni, M. Pepe, P. Tagliolato, and P. Carrara. EDI - a template-driven metadata editor for research data. Journal of Open Research Software, 4a, 2016.

[xii] P. Martin *et al.*, "Open Information Linking for Environmental Research Infrastructures," *2015 IEEE 11th International Conference on e-Science*, Munich, 2015, pp. 513-520. doi: 10.1109/eScience.2015.66