LifeWatch
e-Science European Infrastructure
for Biodiversity and Ecosystem Research

Thesauri & Semantics in the Ecological Domain

Report

Lecce, 9 - 10 June 2016

Authors

Caterina Bergami, CNR – IBAF
Nicola Fiore, University of Salento
Alessandro Oggioni, CNR – IREA
Ilaria Rosati, CNR – IBAF
Paolo Tagliolato, CNR - ISMAR

http://www.lifewatchitaly.eu

Table of Contents

Introduction

Recently, one of the major challenges in the ecological science has been managing and searching an increasingly large volume of data, collected in relatively short time periods and across multiple disciplines. Many different standards, approaches, and semantic tools have been developed to improve data sharing and interoperability. Modern semantic technologies provide a promising way to properly describe and interrelate these different data sources in ways that reduce barriers to data discovery, integration, and exchange among ecological resources and researchers.

Several vocabularies have been already produced in the environmental and ecological domain, such as Earth (Environmental Application Reference Thesaurus[1]), EnvThes (Environmental Thesaurus[2]) and BODC (British Oceanographic Data Centre[3]), just to name a few.

Also, the e-Biodiversity Research Institute of LifeWatch-Italy, which supports scientific research on biodiversity, its relationships with ecosystems services and societal benefits, is trying to tackle this challenge through the development of thesauri and ontologies, concerning overall functional traits. However, the efficient use of these resources is still prevented by the lack of a standardised framework for their mapping which could allow using data described by different vocabularies in a transparent, disciplinary independent, and scientifically valid way.

On 9 and 10 June, LifeWatch-Italy, the Italian distributed centre of LifeWatch, the e-Science European Infrastructure for Biodiversity and Ecosystem Research, held the workshop “Thesauri & Semantics in the Ecological Domain” at the University of Salento (Lecce, Italy). The Workshop engaged 20 participants representing 8 organisations, active in the development and application of semantic tools in the ecological domain (the complete participant list is included at the end of this report).

The Workshop aimed at:

- Sharing good management of semantics practices in the context of ecology;
- Discussing collaboration opportunities in the development and integration of thesauri/vocabularies.

The first day of the workshop saw a series of presentations made by participants, with different scientific and technical expertise, to share current practices, tools, governance, also aiming at enhancing the relations within the groups working on ecological thesauri. The second day, three discussion sessions were

[1] http://thesaurusonline.iia.cnr.it/tematres/earth/
[2] http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn
[3] http://www.bodc.ac.uk

held 1) structuring ecological thesauri and semantic sources, 2) from vocabularies to ontologies, and 3) opportunities for collaboration in new proposals.

Presentations included:

Introductory presentation

Alberto Basset
(LifeWatch Italy - JRU Manager)
Introduction to LifeWatch-Italy and the Workshop activities

EnvThes - status, plans and vision

Barbara Magagna
(Umweltbundesamt, Environment Agency Austria)
The Environmental Thesaurus for Long Term Ecosystem Research: technical approaches, contents and on-going activities

BODC Vocabularies in the Oceanographic domain

Alexandra Kokkinaki
(British Oceanographic Data Centre)
NERC vocabularies: features, application use, technologies and governance

Developing the multilingual aspects of EnvThes, a Thesaurus for Ecosystem Research and Environmental Sciences

Nicolas Bertrand
(Centre for Ecology and Hydrology, UK)
A strategy to improve multilingual support:
– Governance - Language expert groups
–Automated suggestion of terms in different languages

LifeWatch-Italy Thesauri: methodological approach, use cases, future developments

Ilaria Rosati
(CNR-IBAF; LifeWatch-Italy)
LifeWatch Thesauri, the methodology used for their provision, their utility in discovering and integrating multiple data

The RITMARE SDI: from a national project to the Web of Data

Cristiano Fugazza
(CNR-IREA, Italy)
How INSPIRE can evolve to the Web of Data

Geographic names and "Vocabularies". Current developments in LifeWatch-Italy

Paolo Tagliolato
(CNR – IREA; LifeWatch-Italy)
Managing geographic features as semantic resources: motivations and goals. Report on recent activities in LifeWatch Italy: modelling IGM Italian toponyms according to the Geonames Ontology.

LifeWatch-Greece data-services: on supporting metadata and semantic integration for the biodiversity domain

Nikos Minadakis
(FORTH; LifeWatch-Greece)
The approach of LifeWatch-Greece to:
- Support cataloguing and publishing of all the relevant meta-data information of the biodiversity domain;
- Integrate data from heterogeneous sources;
- Efficiently discover biodiversity data.

Outputs and follow up activities

The variety and richness of discussion at the Workshop aimed at drawing together several insights, observations and proposals for future actions. Overall, the Workshop provided a very good opportunity to exchange information about management practices of semantics, both thesauri and ontologies, in the context of ecology. There were also opportunities to exchange practical experiences with the speakers and among participants.

In particular, the different groups shared and discussed the current practices for the development and implementation of thesauri, and decided to produce two publications from the outputs and discussions of the workshop. Specific information on these is given in the part of the report dedicated to the two round tables.

During the discussion, participants also presented the governance model of their organisations in the management of thesauri (presence of owners, managers, publishers and steering committees), and the procedures aimed at tracking changes in the structures or terms of the different thesauri. Different kind of governance used by different groups will be described and synthetised in a schema which can be useful for future publications.

The groups may try to converge in terms of workflow tracking procedures and governance, if it would not be possible to converge towards a single schema, participants will try to propose two different ones to be discussed and possibly shared in the future.

A virtual meeting will be planned for the end of July to further discuss best and common practices for the management and implementation of thesauri (with a focus also on mapping tools and procedures).

The two round tables provided the ideal forum for the final debate and formulation of recommendations regarding the activities to be undertaken for further progress.

The workshop will be used to jointly develop proposals in different context: the Horizon2020 programme at the European level (Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation) and the Pre Commercial Procurement Calls at the national/regional level.

The conclusions and discussion drawn from the Workshop can be further deepened thanks to the dedicated website (http://www.servicecentrelifewatch.eu/web/lifewatch-italia/thesauri-semantics-in-the-ecological-domain), where more detailed information on the proceedings are available, including videos of all the presentations. The expectation is that individuals and organisations will have been motivated by these discussions to develop and implement more shared management practices of semantics in different domain of knowledge.

Round Table - Structuring ecological thesauri and semantic sources. A working group session to discuss current practices, tools and possible relations with the Observation and Measurements paradigm

The aim of this round table was to discuss the current practices within those groups working on environmental thesauri. Semantic technologies, SKOS thesauri, references to authoritative sources for annotating environmental and ecological data, are factors fostering metadata interoperability. The opportunity to discuss together the single practices and, possibly, share the experiences could constitute the background for further developments in this direction.

During the round table several topics were discussed. These are listed below and some notes on the outcomes have been added.

Technologies:

- Tools for thesauri composition: tools' characteristics and considerations on their effectiveness in supporting the working community.
    - Different experiences characterise the work of the main three communities involved in the Workshop, namely NERC (NVS2 Vocabulary Server), LTER EnvThes and LifeWatch-Italy thesauri initiative.
        - NERC developed its own software for the management of vocabularies, so the tools perfectly fit the user requirements. The software is not open source and, concerning the backend, it exploits both an Oracle database and a Fuseki[4] (Apache, open source) SPARQL 1.1 endpoint. Versioning takes place and the recommended practice for old versions is deprecation (not deletion).
        - EnvThes based its work on the "TopBraid" solution. The software exploits Fuseki SPARQL 1.1 endpoint as well. The tool enables extra *rdf* annotations, so that notes can be shared among editors. The tool also supports "working copies", in which the editors can work before publication. In a way, in this case, the tool drives the work of the community. In order to accomplish the entire workflow task, other tools are used in parallel (e.g. google documents, spread sheets).
        - LifeWatch-Italy thesauri were developed with TemaTres open source software. It comprises a SPARQL 1.0 endpoint and a relational database. LifeWatch Service Centre corrected and customised some functionalities of the software, but the current idea is to migrate towards a more robust system. Communication among editors is conducted by the "private note" in the thesaurus working copy, emails

[4] https://jena.apache.org/documentation/serving_data/

and physical and/or web meetings (the community is organised in very small communities of experts working on very specific vocabularies forming the corpus of the whole initiative).

- SPARQL versions. Does your practice involve federated queries and in which cases you make use of them?
    - It emerges as, particularly for mapping among the different initiatives, the preference for v 1.1 of SPARQL. This version favorites federated queries, even if not all the current practices involved them.
- Mapping vocabularies.
    - Different tools for composing mappings have been reported by participants. Among them Silk[5] (EnvThes, LifeWatch), are under current consideration and testing. LifeWatch-Greece had previous experience with this tool and pointed the attention on some issues about its scalability, suggesting the command line tool “X3ML”currently in use by their group.
- Reasoning and SPARQL. (Do you make use of reasoners within your endpoints, e.g. for resolving *owl:sameAs* predicates? Would you suggest any scenario involving reasoning?)
    - There is still a lack of experience when we address reasoning applied to thesauri. One point that was discussed concerned the possibility of enabling simple reasoning on *owl:sameAs* predicates in a federated scenario, where mappings could be traced and equivalences among concepts could be automatically resolved.

Conceptual organisation:

- SKOS/RDF Practices for grouping vocabularies concepts (e.g. named graphs, collections, separate thesauri, etc.). What kind of documentation do you offer with respect to this organization (e.g. external documents, *rdfs:label*, etc.)?
    - Website documentation is the main source of information for the communities involved. Moreover, where communities developed their own tools, their web interfaces suggest such informative layer to users.
- Use of extra-SKOS schemas (RDFS, SKOS-XL, etc.)
    - The usage of SKOS-XL in AGROVOC was reported, but no participant had direct experience in that direction.
- Use and structure of thesauri in relation to Observations and Measurements

[5] http://silkframework.org

- Observations and Measurements conceptual model is taken into consideration in the internal organisation of NERC vocabularies. Within NERC, this conceptual model, is exploited for specific parts (e.g. observed properties); EnvThes is a whole thesaurus whose top concepts are directly inspired by O&M concepts. During the discussion a different reference model emerged by LifeWatch-Greece, namely CRMSci the Scientific Observation Model, inspiring the work with ontologies of the group. Within LifeWatch-Italy the discussion on O&M is in progress, but currently the organization of their thesauri is not reflecting the model. The emergence of OM-lite, a recent ontology from CSIRO, was reported by NERC, when discussing the opportunity to trace relations among concepts (e.g. admissible dimensions and units of measure for observable properties of a given feature of interest type). A proposal was to map this kind of relations in order to enable a finer support to tools exploiting semantic resources for O&M document composition.

Participants agreed on the will to continue the discussion in next online and live meetings, and cooperate in the production of a common report as starting point for further collaborations.

Round Table - From vocabularies to ontologies. A working group session to discuss how a semantic approach could be useful for the discover, search, interoperability and analysis of biological data.

The focus of the round table was to discuss which is the role of vocabularies and what are their links with ontologies, how a semantic approach in the biological field is useful for data discovery and integration, and start to investigate if it is possible to test the benefit of these technologies in data quality and analysis processes.

During the Workshop the following semantic approaches have been introduced:

- ✓ **LifeWatch-Italy Core Ontology**, that is an extension of the OBOE model combined with the work done with the thesauri and it is the actual model used in the LifeWatch-Italy architecture to map all the datasets supplied by the data provider. LifeWatch-Italy team is working on the design and development of new tools that will allow researchers not only to discover data, but also support them in the data analysis.

- ✓ **CIDOC CRM Extensions** (CRMSci, CRMgeo, MarineTLO) is the approach adopted by the LifeWatch-Greece team. CIDOC CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relations used in the cultural heritage domain. This can be applied effectively to a variety of domains such us biodiversity, geology, etc. (ISO 21127:2006). CRMsci is a formal ontology indented to be used as a global schema for integrating metadata about scientific observation, measurements and processed data in a descriptive and empirical way. MarineTLO aims at being a global core model that provides a common and agreed-upon understanding of the concepts and relations holding in the marine domain, so to enable knowledge sharing, information exchanging and integration among heterogeneous sources, covering with suitable abstractions the marine and the terrestrial domains to enable the most fundamental queries.

- ✓ **OM Lite** is an OWL ontology for observations and sampling features, based on the O&M conceptual model from ISO 19156. It is the approach adopted by the NOC (National Biogeographic Centre).

The discussion focused on which are the potentialities of all these models, if they could be used in a proficient way not only to discover data, but also to assure their quality, and if these approaches could be a useful support in data analysis.

In order to evaluate all these aspects, participants agreed to start an evaluation phase testing all the models on a specific showcase based on the phytoplankton domain. The proposed working plan that includes the following phases: providing data (along with metadata), defining questions,

mapping with OM Lite, CRM EXT, LW-ITA Core Ontology. The partners involved are LW-ITA, BODC, LW-GR, LTER, CEH.

Participants agreed also that at the end of the experimentation phase, they will collaborate in writing a paper presenting the obtained results.

It is also important to highlight that in the future, it will be important also to work in a way to make the different models interoperable.

List of participants

Alberto Basset, University of Salento

Alessandro Oggioni, CNR - IREA

Alexandra Kokkinaki, British Oceanographic Data Centre

Angela Boggero, CNR - ISE

Anna Gazda, University of Agriculture in Krakow

Barbara Magagna, Umweltbundesamt

Cataldo Pierri, CNR - IBAF

Caterina Bergami, CNR - IBAF

Cristiano Fugazza, CNR - IREA

Elena Stanca, University of Salento

Ilaria Rosati, CNR - IBAF

Mario Bochicchio, University of Salento

Nicola Fiore, University of Salento

Nicolas Bertrand, Centre for Ecology and Hydrology

Nikos Minadakis, FORTH

Paolo Plini, CNR - IIA

Paolo Tagliolato, CNR - ISMAR

Quentin Groom, Botanic Garden Meise

Sham Navathe, Georgia Institute of Technology

Stefano De Felici, CNR - IBAF