Metadata, Controlled Vocabularies and Ontologies 2nd Working meeting Report

Lecce, 15 -17 April 2019

Authors

Antonio José Sáenz, LifeWatch ERIC

Alessandro Oggioni, LifeWatch Italy (CNR – IREA)

Caterina Bergami, LifeWatch Italy (CNR – ISMAR)

Ilaria Rosati, LifeWatch Italy (CNR – IRET)

José María García, LifeWatch Spain (University of Seville)

Magda Aljancic, LifeWatch Slovenia (Karst Research Institute)

Nicola Fiore, LifeWatch ERIC

Zhiming Zhao, LifeWatch Netherlands (University of Amsterdam)

Executive Summary

The LifeWatch ERIC (hereafter LW ERIC) Service Centre organised the second working meeting on “Metadata, Controlled Vocabularies and Ontologies” in Lecce from 15 to 17 April 2019.

The aim of the meeting was to continue the activities related to the definition of 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 10 experts, with both scientific and technical backgrounds, from 5 national nodes of LW ERIC (Greece, Italy, Spain, Slovenia, The Netherlands). Three sessions were held on 1) Metadata, 2) Controlled Vocabularies, and 3) Ontologies, each one including a summary on work already done and working sessions for the implementation and curation of metadata, standardized controlled vocabularies and ontologies.

Table of Contents

Ontologies

Discussion & Working session

Phytoplankton Show Case Requirements Elicitation

In the introduction session, a summary of the first face-to-face meeting and regular calls has been made in order to define the main goals of this meeting and the next steps necessary in order to implement the LW ERIC ontology.

The team agreed to adopt a bottom-up approach to create the ontological model and to start with the Phytoplankton showcase thanks to the contribution of domain experts from the LifeWatch Italy Node. The work will be carried out considering both technical and conceptual aspects.

To achieve these objectives, adopting an engineering approach on the specific domain (following ENVRI+ best practices and reference model), the team began to analyze the scientific methodology adopted in this showcase, in order to understand the life cycle of the data.

Using the ENVRI reference model as a guide (https://confluence.egi.eu/display/EC/Model+Overview) the following main phases were identified and described.

FIRST PHASE

DATA ACQUISITION (loose term)

- SAMPLING DESIGN*

  The definition of the spatial and temporal scale of the sampling. This could be also accordingly with different national or international Directives (e.g.: WFD, MSFD) or with specific project requirements;

- SAMPLING PROTOCOL**

  Specific methodologies to be used when sampling (e.g.: samples collector);

- SAMPLES ANALYSIS and digitalization

  Specific methodologies to be used when analysing the samples (e.g.: qualitative or quantitative analyses using a image analysis software, NIS/NIKON or LUCIA/NIKON), taxonomic identification, species-specific linear dimensions measurement, excel file population according to the LW Data template.

*ENVRI Reference: “Data handled at this phase include raw data products, metadata and processed data. Where possible, processed data should be reproducible by executing the same process on the same source data-sets, supported by provenance data. Operations such as data quality verification, identification, annotation, cataloguing, replication and archival are often provided. Access to curated data from outside the infrastructure is brokered through independent data access mechanisms. There is usually an emphasis on non-functional requirements for data curation satisfying availability, reliability, utility, throughput, responsiveness, security and scalability criteria.”

** Actually, in these processes all the information is in free text (using natural language), we need to structure them to be machine-readable. It is important also to define a domain language.

The Output of this phase is: DATASET (RAW DATA) https://drive.google.com/open?id=15AEto9C_MfA-rFf-njBHm3c9ud70xYd1

SECOND PHASE

- DATA CURATION

  There is the need to i) assign PID to the dataset in the different states (raw data, intermediate, published data) and ii) provide rich contextual meta information to those states.
  It is necessary also to take in account the provenance aspects, and the LifeBlock Services that are pervasive to the whole e-infrastructure.

- DATA ANALYSIS

  Calculation of other morphological traits (Biovolume, Biomass, Carbon Content, etc.) starting from Linear Dimensions; computation of the indexes; distribution analysis on morphological and demographic traits and indexes.

- RESULTS PUBLICATION

  From LifeWatch point of view is important to make the following assets FAIR:

  - the DATASETS in all the states
  - the Papers, Reports, Software, Research Objects, etc...
  - all the scientific results

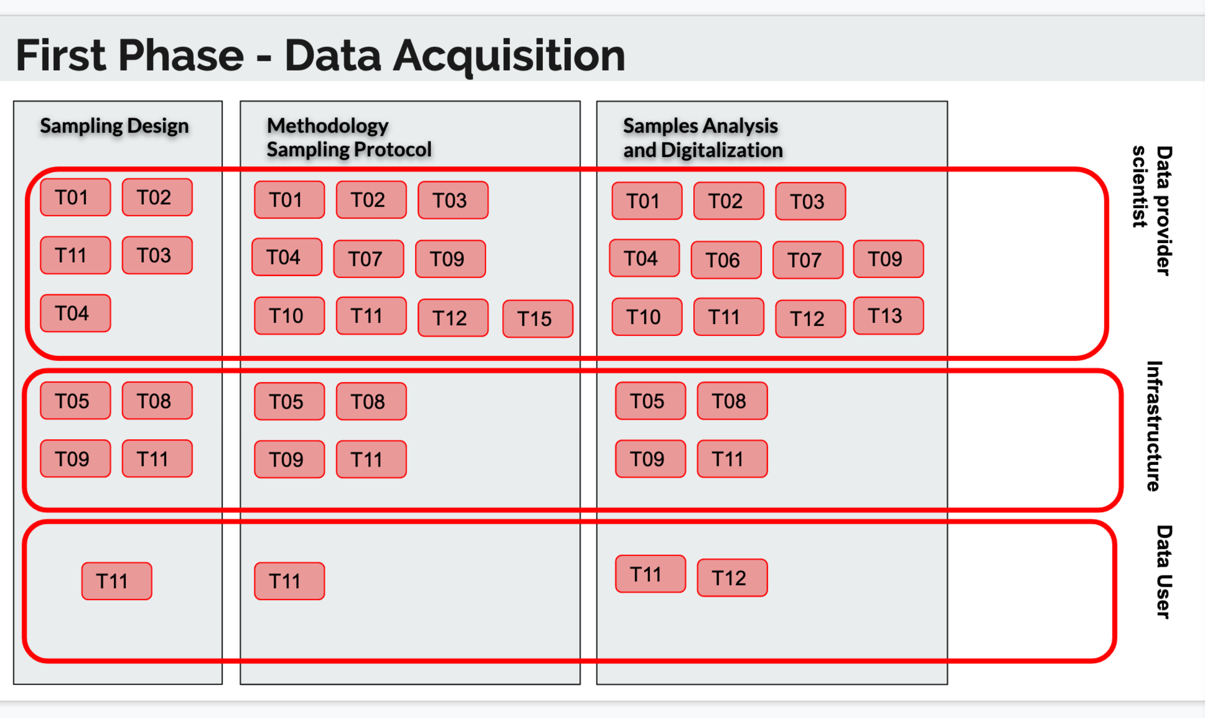
Starting from this requirements elicitation, the participants worked to identify which tools and software can support these phases. In this process, the Requirement List (https://docs.google.com/document/d/1s0qhK4KfyTOR1WXE3AyfgpVX9cXKSAKl2R-

KOMSpQnY/edit) was used.

We identified the following list of potential tools:

1. DATA CURATION - Context-Aware Data Curation Tool (Metadata Management)
2. DATA CURATION - Metadata Mapping Tool
3. DATA PUBLISHING - (Meta)Data Catalogue
4. DATA CURATION - PID Service
5. METADATA QUALITY CHECKER
6. QUALITY ASSURANCE PUBLISHER SERVICE (merged with 3)
7. CURATION: LONG TERM PRESERVATION SERVICE (user/data provider customized Data management plan + Service Level Agreement + other issues)
8. ACCOUNTING SERVICE FOR LIFEWATCH ASSETS
9. PROVENANCE SERVICE
10. DATA SUBMISSION SERVICE (with metadata included)
11. SEMANTIC DISCOVERY/SEARCH TOOL
12. DATA ACCESS TOOL
13. CONSISTENCY CHECKER TOOL
14. WORKFLOW MANAGEMENT TOOL
15. Workflow process modeling and metadata description

All the tools should provide not only GUI interfaces but also API interfaces. Furthermore, the team assigned the different tools to the identified steps of the first phase of the Phytoplankton Showcase, considering also the different users (Data provider scientists, Infrastructure, Data users).



First Phase - Data Acquisition

	Sampling Design	Methodology Sampling Protocol	Samples Analysis and Digitalization
Data provider scientist	T01 T02 T11 T03 T04	T01 T02 T03 T04 T07 T09 T10 T11 T12 T15	T01 T02 T03 T04 T06 T07 T09 T10 T11 T12 T13
Infrastructure	T05 T08 T09 T11	T05 T08 T09 T11	T05 T08 T09 T11
Data User	T11	T11	T11 T12

FIRST PHASE: Data acquisition

- SAMPLING DESIGN

The definition of the spatial and temporal scale of the sampling. This could be also accordingly with different national or international Directives (e.g.: WFD, MSFD) or with specific project requirements.

Data provider/scientists

- Tool 1: contextual curation. (see Req. 1)
- Tool 2: mapping tool (browse/select/search mappers, create mapper, apply the mapper to transform)
- Tool 11: check existing work (search, viewers, store, notes)
- Tool 3: publishing (publishing request form)
- Tool 4: PID service (provide metadata, request PID, be invoked by publishing service)

Infrastructure:

- Tool 5: check the quality of metadata (automatically tricked by human changes, e.g., upload/edit/delete, and user-friendly feedback)
- Tool 9: provenance (automated tool, query/visualization, configure)

- Tool 11: check existing work of the other users,
- Tool 8: accounting service (Automated triggered, query/visualization/ configuration)

Data user:

- Tool 11: check existing work

- Methodology: SAMPLING PROTOCOL

Specific methodologies to be used when sampling (e.g.: samples collector)

Data provider/scientists/

- Tool 1: contextual curation.
- Tool 2: mapping tool
- Tool 11: check existing work
- Tool 3: publishing service
- Tool 4: PID service (provide metadata, request PID, be invoked by publishing service,)
- Tool 10: submission tool (invoke tool 1, submission form/gui)
- Tool 7: preservation tool (describe DMP, negotiation of SLA)
- Tool 9: provenance
- Tool 12: data access (retrieve, format transformation, visualization, gui,)
- Tool 15: describing the protocol and steps of scientific processes ()

Infrastructure

- Tool 5: check the quality of metadata
- Tool 9: provenance
- Tool 11: check existing work of the other users,
- Tool 8: accounting service

Data user

- Tool 11: check existing work

- SAMPLES ANALYSIS

Specific methodologies to be used when analysing the samples (e.g.: qualitative or quantitative analyses using an image analysis software, NIS/NIKON or LUCIA/NIKON)

Data provider/scientists/

- Tool 1: contextual curation.
- Tool 2: mapping tool
- Tool 11: check existing work
- Tool 3: publishing service
- Tool 4: PID service
- Tool 10: submission tool
- Tool 7: preservation tool
- Tool 9: provenance
- Tool 12: data access
- Tool 13: consistency check

Infrastructure

- Tool 5: check the quality of metadata
- Tool 9: provenance
- Tool 11: check existing work of the other users
- Tool 8: accounting service

Data user

- Tool 11: check existing work
- Tool 13: consistency check (input: PID/metadata/list of repositories)

All the work done allow the team to define a desirable ROADMAP that LifeWatch ERIC could follow in the mid term period.

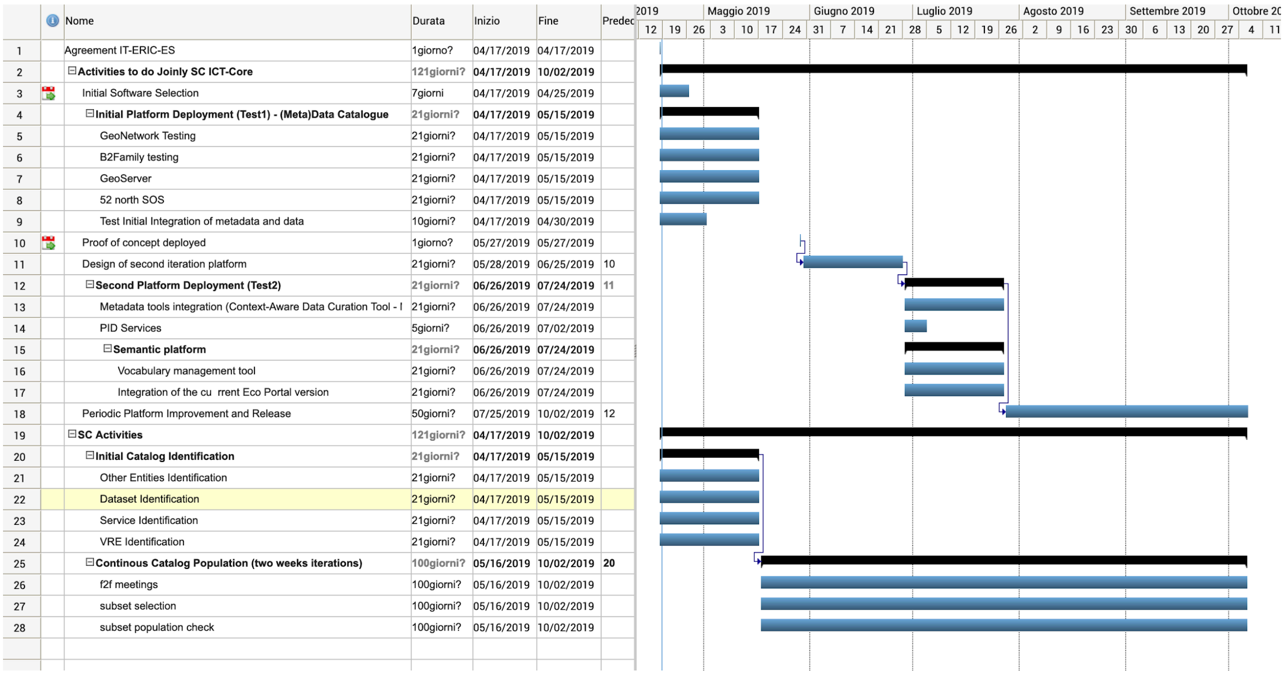
ROADMAP

PRIORITY	ID	TOOL	SUGGESTED SOFTWARE[1]
1	1	DATA CURATION - Context-Aware Data Curation Tool (Metadata Management)	EDI
1	2	DATA CURATION - Metadata Mapping Tool	Hale Studio and GeoDab
1	3	DATA PUBLISHING - (Meta)Data Catalogue	GeoNetwork and B2Share
1	4	DATA CURATION - PID Service	B2Handle
2	5	METADATA QUALITY CHECKER	to be defined
2	10	DATA SUBMISSION SERVICE (with metadata included)	to be defined
2	11	SEMANTIC DISCOVERY/SEARCH TOOL	B2Find
3	8	ACCOUNTING SERVICE FOR LIFEWATCH ASSETS	It is being analized and defined by ICT-Core
3	9	PROVENANCE SERVICE	LifeBlock (LifeWatch blockchain distributed infrastructure) Provenance

[1] link is to the GitHub repository, if the software is open source, otherwise the link is to the website

			Template (developed in theme2)
3	12	DATA ACCESS TOOL	to be defined
3	13	CONSISTENCY CHECKER TOOL	to be defined
3	14	WORKFLOW MANAGEMENT TOOL	to be defined
3	15	Workflow process modeling and metadata description	to be defined

On the base of the roadmap, a first Draft of the Working Plan has been defined, identifying deadlines and responsibilities (Nicola Fiore as Service Center and Antonio José Sáenz as ICT-Core, see below the table).



		Nome	Durata	Inizio	Fine	Predec
1		Agreement IT-ERIC-ES	1giorno?	04/17/2019	04/17/2019	
2		⊟**Activities to do Joinly SC ICT-Core**	121giorni?	**04/17/2019**	**10/02/2019**	
3	📅	Initial Software Selection	7giorni	04/17/2019	04/25/2019	
4		⊟**Initial Platform Deployment (Test1) - (Meta)Data Catalogue**	21giorni?	**04/17/2019**	**05/15/2019**	
5		GeoNetwork Testing	21giorni?	04/17/2019	05/15/2019	
6		B2Family testing	21giorni?	04/17/2019	05/15/2019	
7		GeoServer	21giorni?	04/17/2019	05/15/2019	
8		52 north SOS	21giorni?	04/17/2019	05/15/2019	
9		Test Initial Integration of metadata and data	10giorni?	04/17/2019	04/30/2019	
10	📅	Proof of concept deployed	1giorno?	05/27/2019	05/27/2019	
11		Design of second iteration platform	21giorni?	05/28/2019	06/25/2019	10
12		⊟**Second Platform Deployment (Test2)**	21giorni?	**06/26/2019**	**07/24/2019**	11
13		Metadata tools integration (Context-Aware Data Curation Tool - I	21giorni?	06/26/2019	07/24/2019	
14		PID Services	5giorni?	06/26/2019	07/02/2019	
15		⊟**Semantic platform**	21giorni?	**06/26/2019**	**07/24/2019**	
16		Vocabulary management tool	21giorni?	06/26/2019	07/24/2019	
17		Integration of the cu rrent Eco Portal version	21giorni?	06/26/2019	07/24/2019	
18		Periodic Platform Improvement and Release	50giorni?	07/25/2019	10/02/2019	12
19		⊟**SC Activities**	121giorni?	**04/17/2019**	**10/02/2019**	
20		⊟**Initial Catalog Identification**	21giorni?	**04/17/2019**	**05/15/2019**	
21		Other Entities Identification	21giorni?	04/17/2019	05/15/2019	
22		Dataset Identification	21giorni?	04/17/2019	05/15/2019	
23		Service Identification	21giorni?	04/17/2019	05/15/2019	
24		VRE Identification	21giorni?	04/17/2019	05/15/2019	
25		⊟**Continous Catalog Population (two weeks iterations)**	100giorni?	**05/16/2019**	**10/02/2019**	**20**
26		f2f meetings	100giorni?	05/16/2019	10/02/2019	
27		subset selection	100giorni?	05/16/2019	10/02/2019	
28		subset population check	100giorni?	05/16/2019	10/02/2019	

Metadata

Discussion & Working session

The discussion on metadata has been moved into the debate on Ontology and Phytoplankton showcase. The team involved in the discussion started from the “Metadata Requirements for LifeWatch ERIC” document, provided to all participants before the meeting, in order to define a list of requirements for the metadata management in LifeWatch ERIC. The team discussed the necessary requirements for the tools/softwares that could better meet the needs of the LifeWatch ERIC. Team recognised 3 general requirements:

1. All the assets provided by those tools could be shared with group members or with social networks (e.g. twitter, facebook, google+, email);
2. Geoinformation quick viewer/selection;
3. Personal context (e.g. searching, selection, etc.) should be stored (e.g. save bookmarks, etc.).

The team also identify as an important issue to distinguish between the requirements needed for the **Data Owners/Providers** and the ones for the **Data Users/Consumers**.

Data Owners/Providers

1. A dynamic and user-friendly metadata management interface to define different metadata sets for all the selected entities;
2. Dynamic and user-friendly metadata entry tool with the following functionalities:
    a. autocompletion functionalities based on the triple store;
    b. the possibility to explore all tabs and editing the record as necessary;
    c. define a custom metadata schema profile (e.g. multiplicity, mandatory/optional);
    d. insertion of references to semantic resources (URIs of RDF sources with linked data/sparql) from corresponding text labels presented in the user interface;
    e. insertion of references to publications (DOI), people (ORCID) or PIDs related to other entities;
    f. insertion/display of licenses for use of entities;
    g. update, versioning and management metadata records;
3. A module to map the external metadata on LW used metadata schemas:
    a. visualize current mappers, standards, formats etc.
    b. create new mappers

   c. use mappers to creates new records/formats;

4. A module to harvest the mapped and translate metadata sets;

5. A module to assign a persistent identifier (PID) to all entities;

6. A module for providing data usage statistics (metrics such as metadata viewed, data viewed, data downloaded, data cited, etc.);

7. A module to export metadata according to different schemas and formats;

8. A module for metadata quality control, detects and corrects (or removes) corrupt, inconsistent or inaccurate records from data sets;

9. A module for publishing curated data / metadata;

10. A module for depositing (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and makes them accessible on request- Data management plan, and Service Level Agreements (SLA).

Data Users/Consumers:

1. Provide a range of query interfaces to accommodate various data search behaviors (Simple, Advanced and Map search);

2. Provide multiple access points to find data (e.g. search, subject browse, faceted browse/filtering). Facets are usually derived from controlled vocabularies (e.g. subject, data type, file format, etc). Data repositories and data providers should work together and adopt community accepted vocabularies, this will give users a consistent search experience across repositories;

3. Make it easier for researchers to judge the relevance, accessibility and reusability of a data collection from a search summary (e.g. to make it clear if data are accessible; to make the data license clear; etc.);

4. Expose data usage statistics (metrics such as metadata viewed, data viewed, data downloaded, data cited, etc.);

5. Strive for consistency with other repositories;

6. Identify and aggregate metadata records that describe the same data object;

7. Make metadata records easily indexed and searchable by major web search engines;

8. Follow API search standards and community adopted vocabularies for interoperability;

9. A module/tool for user to seamlessly access data, using (aggregated) metadata to retrieve data sets from different locations, and create contextual metadata for workflow to further process.

Finally, the team agreed on the list of entities to be considered in LifeWatch ERIC: Dataset, Network, Site, Station, Instrument, Publication, People, Activity, Programme/Project, Sample, Infrastructure, Sampling protocol/Method, Sampling design, Service, VRE, and Workflow. During the discussion, the team provided for the most of entities the definition, xsd schema and link to the documentation.

Controlled Vocabularies

Discussion & Working session

During the last month (starting from the mid of March 2019), the working group tested the five tools for the implementation, curation and publication of controlled vocabularies, selected during the previous meeting in November 2018. The delay in starting this activity was due to the approval delay from the Executive Board of the LW ERIC of the first report (1st february 2019), in which all the future activities and strategies identified during the november meeting were listed. Moreover, the VPN connection and the first testing tool (TemaTres) were available for the WG in mid March.

Since now, the tested tools are: TemaTres, Themas, Thesauform, and VocBench. The working group tested them based on the list of functionalities recognised during the previous meeting and divided between requirements for thesauri implementation and curation and thesauri publication. The list of the requirements and the synthetic results and comments from the test phase are collected in this file: https://docs.google.com/spreadsheets/d/1okGbdLP6p8d4E0ZlJlOb-wXf2QvvoikhNM2JF-aeJYA/edit?usp=sharing.

In the following section, we describe in details the results of all the tests.

We started testing **TemaTres**, which is the tool used by LifeWatch ITA in the last years for the implementation of its thesauri. We testeds the last released version, 3.1, which shows new functionalities with respect to the oldest ones, such as the improved collaborative workflow, where now it is possible to trace the different type of notes and also the editors' name. Moreover, the tool in its new version enables a multilingual editor facility. Infact, by clicking on the multilingual editor, is possible to create a relationship (equivalent, partial equivalent, not equivalent) between the selected term and a term in a target vocabulary. The result in SKOS is an exact match between the two terms, instead of the creation of another prefered Label for the same concept with an attribute specifying the language (e.g.: xml:lang="it"). In this way the term is not simply translated but is in relation with another new term (the relationship is visible only in the selected term and not in the target one). However, in this new version, this tool also has a number of gaps in its functionalities, such as no triggering capabilities, no possibility of having multiple projects and thesaurus versioning.

The second tested tool was **Themas**, a tool developed and used within DARIAH European Research Infrastructure and proposed by LifeWatch Greece. Themas covers a wide range of functions such as a collaborative workflow via user Roles with different rights, the possibility to have multiple projects, procedures for the mass import/export of terms in different formats, alternative forms of navigation, the support of complex search criteria and scalability search. The tool can also create and manage multilingual thesauri with Greek or English language serving as the dominant language and a configurable number of reference translation languages but we were not able to evaluate this aspect since this functionality did not work during the tests. **Themas tool** is not able to create and manage notes such as definition, bibliographic note and scope note and it has a basic versioning. Moreover, this tool does not present important functionalities related with the vocabulary publication as web interface and sparql endpoint.

Regarding **Thesauform** tool**,** it was developed by the Centre for the Synthesis and Analysis of Biodiversity (CESAB; French Foundation for Research on Biodiversity - FRB) with the aim to create a web based tool dedicated to the collaborative construction of a thesaurus by experts in the field of plant functional diversity research. Thesauform has three components: contribution, validation, and publication. The contribution component has the annotation module, a tool for the domain expert to contribute and comment on terms and their definitions relevant for their community; and the voting module (dropdown menu with 6 voting categories), an opportunity for the proposed terms and their definitions to be evaluated by the specific domain community. The validation component is a unique attribute of Thesauform, that allows an editorial team to critically review the contributions and determine the best product to be published. This component also contains the administration of the site. The administrators can upload and download bulk data for the contributors and manage access. Only administrators can enter information using the ‘Administration’ tab. Once a term and its associated properties has been validated, they can published it. All terms can be discovered through different searches capabilities, and the API is also provided. Thesauform shows very user-friendly functionalities related to creation and validation of terms, however, the tool has a number of gaps, such as no triggering capabilities, no possibility of having multiple projects, no multilingual editing and only basic versioning.

At last we tested was **VocBench 3,** which is a web-based, multilingual, collaborative development platform for managing OWL ontologies, SKOS(/XL) thesauri, Ontolex-lemon lexicons and generic RDF datasets. It was realized by the ART Research Group at the University of Rome Tor Vergata and used by the Food and Agriculture Organization (FAO) for producing AGROVOC, a controlled vocabulary consisting of more of 36000 concepts, available in more than 33 languages and covering different areas of interest including food, nutrition, agriculture, fisheries, forestry and environment. From our test, VocBench covers a wide range of desired requirements:

1. a very functional collaborative workflow including the possibility to create new roles and groups with different capabilities assigned from administrator and a validation workflow for validating terms, definition and alignments;

2. a multilingual editor facility to create prefered labels in different languages for the

same concept with an attribute specifying the language (e.g.: xml:lang=”it”);

3. a multi projects editor;
4. different import procedures and models;
5. access control and user management;
6. Role Based Access Control (RBAC): it is possible to customize all the roles and capabilities, and even easily create new ones;
7. versioning of concepts, skosnote, scopenote and relationships. It is possible to have also versioning of the whole thesaurus;
8. provenance, everything is stored in RDF using widespread vocabularies for provenance (e.g. PROV-O).
9. mapping and alignment functionality with specific tools for the alignment and its validation;
10. SPARQL support with syntax highlight and completion fed by the vocabularies imported in the managed dataset, support for storable queries/updates (to be verified;
11. API (to be verified);
12. different search capabilities (hierarchical, textbox, custom search, advanced search settings);
13. different export procedures and formats available, it is possible to export also thesaurus, thesaurus configuration and metadata as well as queries (BinaryRDF, JSON-LD, N-Quads, N-Triples, N3, RDF/JSON, RDF/XML, TriG, TriX, Turtle).

At the end of this initial test phase, we decided to adopt VocBench 3 for the implementation of the LW thesauri/controlled vocabularies as it seemed the most suitable tool for the projected work. However, before starting the integration of this tool in LW infrastructure, the ICT group need to verify some technical aspects with the developers of VocBench particularly the ones related with the sparql support and API. Moreover, in the next weeks we will try again to contact the Software Sales Executive of TopBraid EDG in order to have a demo version of the vocabularies management tool to test.