A knowledge base for Exploited Marine Ecosystems

Julien Barde[1], Pascal Cauquil[1], and Norbert Billet[1]

[1] Institut de Recherche pour le Dévelopement, UMR EME 212, Sète, France

Abstract. In 2008, IRD started to work on setting up a knowledge base (named Ecoscope) about Ecosystem Approach to Fisheries domain (EAF) in the context of a marine ecology laboratory studying Exploited Marine Ecosystems in different regions of the world. This application was meant to fit the needs of researchers by improving knowledge and related information resources management [14,12]. Among other goals, researchers expected an information system enabling to provide an inventory of available data sources (ecological observations, satellites images, pictures, articles, reports..) and facilitating data rescue, data access, data processes (indicators..) as well as the ability to summarize related knowledge through fact sheets about domain entities (ecosystems, species..) connected with hyperlinks (based on ecosystem relationships).
Beyond metadata, data management and related interoperability issues (OGC, TDWG...), this project was then a real opportunity to set up an ontology for EAF domain in order to link existing information resources with real-world entities (EAF domain concepts). To achieve these goals, semantic Web standards and reference RDF schemas have been taken into account (SKOS, Dublin Core, FOAF, OBOE, Darwin Core...) and a first version of RDF schema for EAF domain has been set up. These ontologies have been instantiated to describe our information resources and some knowledge about entities that researchers are studying.
A first Website has been set up on top of this knowledge Base. Related Web pages consist mainly in fact sheets about domain entities (ecosystems, top predators and related preys species, fishing vessels..persons) where users can find related information resources (spatial layers, articles, pictures, indicators..). Knowledge can as well be summarized through networks of entities like food webs with dedicated visualizations tools. This is made possible by querying the knowledge base where Linked metadata and data (in underlying databases) are tagged with related species URIs. Proof has been done that Semantic Web languages can be used to fit the needs of our colleagues. Moreover, in the context of iMarine FP7 project, we started to deal with partners having similar projects (FLOD from FAO, Worms, FORTH). We then set up a SPARQL endpoint and OpenSearch access to share the content our knowledge bases with other applications (search engines, text mining applications..).
We will present our current application and related technical choices as well as futur plans to connect additional data sources to enrich this knowledge base and make it available for our partners. In particular we will describe some use cases related to biodiversity management issues.

1 Ecosystem Approach to Fisheries

In this paper, we present our work on knowledge management applied to the domain of Ecosystem Approach to Fisheries (EAF). According to FAO [5], EAF is an approach that:

> strives to balance diverse societal objectives, by taking into account the knowledge and uncertainties about biotic, abiotic, and human components of ecosystems and their interactions and applying an integrated approach to fisheries within ecologically meaningful boundaries.

We used this definition to create a conceptual model as schown in the UML diagram class of Figure 1.

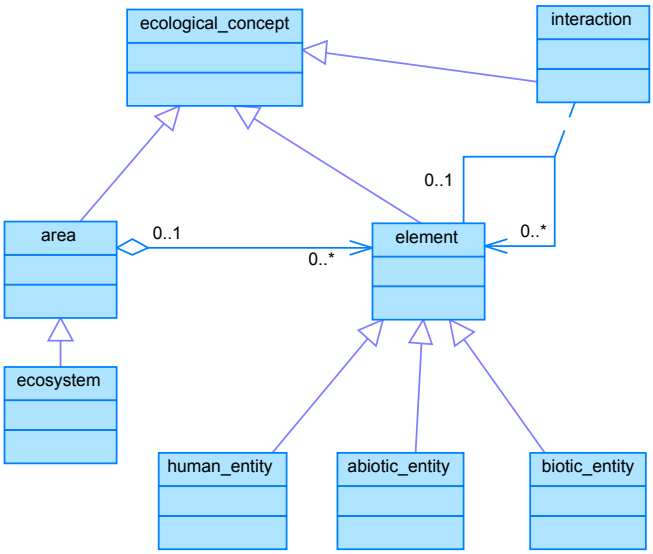

Fig. 1. UML class diagram translating Ecosystem Approach to Fisheries definition [5]

This UML diagram has been used to create a first set of top-level classes and properties for EAF domain (identifying real-world entities according to [2]): ecosystems (areas), their components as well as their interactions. The Figure 2 shows some RDF triples for the class *fish* instantiated for the species *exocoetus volitans*.

Similar objects have been created for hundreds of *species*, *fishing vessels* and *fishing gears*... which are all part of different *marine ecosystems*.

EAF requires thus the management of knowledge related to *marine ecosystems* and their *abiotic*, *biotic* and *human components*. However, this knowledge comes from different scientific studies driven by *researchers* which generate various *information resources*. These entities have to be described as well.

```
   <ecosystems_def:fish rdf:about="&localfile;#exocoetus_volitans">
        <ecosystems_def:wormsId>http://www.marinespecies.org/aphia.php?p=taxdetails&amp;id=126385</
ecosystems_def:wormsId>
        <ecosystems_def:fishbaseId>http://www.fishbase.org/Summary/SpeciesSummary.php?id=1032</
ecosystems_def:fishbaseId>
        <ecosystems_def:faoId>EXV</ecosystems_def:faoId>
        <ecosystems_def:wikiId>http://commons.wikimedia.org/wiki/Exocoetus_volitans</
ecosystems_def:wikiId>
        <skos:prefLabel xml:lang="en">Flying fish</skos:prefLabel>
        <skos:prefLabel xml:lang="fr">Poisson volant</skos:prefLabel>
        <skos:prefLabel xml:lang="es">Pez volador</skos:prefLabel>
        <skos:altLabel xml:lang="fr">exocet</skos:altLabel>
        <skos:altLabel xml:lang="en">Tropical two-wing flyingfish</skos:altLabel>
        <skos:note xml:lang="fr">L'espèce Exocoetus volitans est étudiée au CRHMT en tant que
proie...</skos:note>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans1"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans2"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans3"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans4"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans5"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans6"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans7"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#ale"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#bet"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#swo"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#yft"/>
        <ecosystems_def:used_data_source rdf:resource="&ontologies;/resources#dbStomac"/>
        <ecosystems_def:used_data_source rdf:resource="&ontologies;/resources#dbIsotopes"/>
        <geographic_objects_def:prefGeoGraphicObject rdf:resource="&ontologies;/
geographic_objects#wms_exocoetus_volitans"/>
    </ecosystems_def:fish>
```

Fig. 2. Examples of RDF triples summarizing our knowledge about species

2 Information resources related to EAF

Real-world entities of our domain (like species, fishing vessels, habitats...) are related to different kinds of information resources (pictures, spreadsheets, databases, satellites images, sensors data...)[15,16]. As illustrated in Figure 3, these entities are related as well to some people (agents) who are driving the scientific studies and generating the related information resources by running some processes.

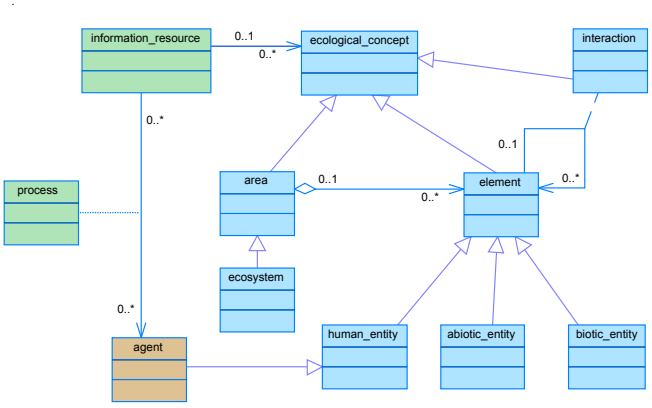

Fig. 3. Information resources, processes and agents related to entities of the domain

As many other laboratories working on ecological studies, the inventory of available data sources consists of various data types (which are subclasses of "information_resource" class in Figure 3):

– ecological observations from fieldwork (observers on-board fishing or scientific vessels collecting samples for data analysis: size, stomach content,

isotopes, contaminants, fatty acids...). These data are usually managed in spatial databases (Postgres / Postgis) or spreadsheets.
- satellites images to characterize environmental parameters of species habitats. These data are usually managed with binary formats (e.g. netCDF files for series of images).
- pictures that are collected by on-board observers or scientists,
- articles, reports... published by researchers,
- processes to run data analysis with different programming languages (R, IDL, Matlab...).

All these resources are usually described and managed with specific (meta-)data formats that impedes basic tasks like data discovery and retrieval. In addition to knowledge management, RDF is expected to facilitate data management (seamless access to metadata catalogues, codelist mapping...) by complying with widely used schemas.

3 Underlying standards

The description and the management of the information resources has to comply with well known standards to ensure that these resources will be made available for various communities of users. In particular, we target communities related to spatial, biodiversity and statistical information. In addition to information resources management, we selected as well existing standards to manage information about agents and domain entities (species...). An effort has been required to transform these (meta)data in RDF with EAF domain URIs.

3.1 Schemas for information resources and related agents

Many standards enable the description of information resources. Most of them consist in XML schema where keywords and mother metadata elements are described with literals (for species, characteristics, fishing vessels, agents... that are observed). This is the case with OGC metadata standards for spatial information (19115, 19119, 19110 [13]...), with metadata standards for biodiversity data (Ecological Metadata language / EML, TDWG standards like Darwin Core [18]...), with SDMX for statistical datasets [17]...

In order to describe and retrieve our information resources by using common metadata elements and URIs we decided to comply with following RDF schemas:

- DCMI [4] as metadata standard which can be used to describe any kind of information resource,
- dclite4g [19] Information model for metadata about geospatial data. ISO 19139 or EML metadata can be converted into dclite4g RDF metadata,
- SKOS [11] for description of terms and definitions related to information resources and concepts,
- FOAF [3] for description of agents (persons, institutes, projects) and their relationships,

– BIBO [7] for bibliographic references.

We aim to describe and make some of our data available as Linked Open Data by taking into account the 5 star development scheme for Open Data [2,1].

However, being able to describe information resources, processes and related agents requires URIs for ecological entities described in Figure 1.

3.2 Standards for domain entities

Among existing RDF schema relevant for our domain, we can mention:

– Previous work on ontologies for ecoinformatics [21].
– OBOE for modeling and representing scientific observations [9,10],
– Darwin Core [20] for sharing of information about biological diversity,
– FLOD [6] for fisheries domain.

These ontologies have been taken into account to map our ontology classes and properties as well as for raw data triplication.

3.3 RDF generation

We have two kinds of RDF triples which are generated from our information resources:

– statements instantiating our ontology with real-world entities (ecosystems and related components): species, fishing vessel, ecosystems... Each data source provides a set of entities (species, environmental parameters...) which are translated into instances of our ontology,
– RDF description of information resources, including related agents. More than basic descriptions, our goal here is to tag metadata with URIs of domain entities (previous item).

For now, we have been using an "ad hoc" approach to get RDF triples from each type of information resource as illustrated in Figure 4. Moreover in some cases, previous efforts can be reused:

– OGC metadata (ISO 19139) can be transformed in GENESI-DEC RDF metadata by reusing an existing XSL file from GENESI-DEC project,
– EML metadata can be transformed in OGC 19139 metadata with a GBIF XSL file,
– bibliographic references metadata can be exported in RDF (BIBO compliant) from Zotero (as well as references of pictures, videos if managed with Zotero).

In this case, the real challenge consists in adding some context to these RDF metadata by relating them to URIs of domain entities. This can be achieved by entity mining approach.

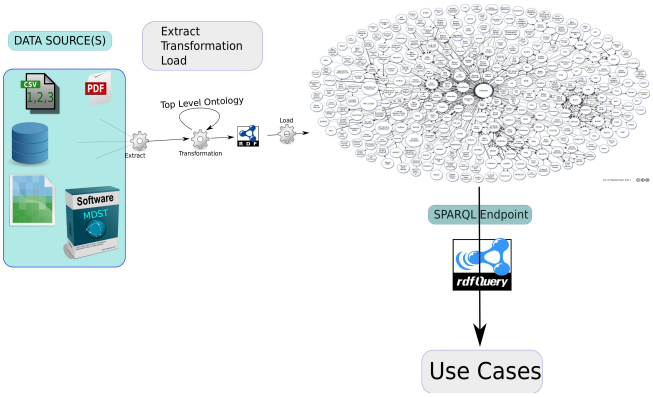

Fig. 4. Data sources to be triplified

4 RDF storage and server

Once classes and properties for EAF entities have been created and related objects (including information resources and agents) instantiated, these objects have been loaded with JENA in a triple store (Jena with TDB, preferred to Postgres for persistent storage and access) and have been made accessible through a SPARQL endpoint (the ontology is available as well at this URL: http://www.ecoscope.org/ontologies/main).

Another access has been set up to deliver search results through OpenSearch protocol with different data formats: HTML, RSS and RDF (Semantic extension).

That was needed to enable a set of use cases and make the content our knowledge base available online.

5 Related applications and products

In this section, we describe some use cases exploiting the content of the Ecoscope knowledge base. Our first use case was the setting up of a Web portal built on top of the ontology / knowledge base (through Jena for storage and access and Struts2 to set up Web pages on top of Jena). The goal was to satisfy basic use cases like data discovery and retrieval, knowledge summary about domain entities (species, fishing vessels. . .).

5.1 Metadata catalogue

One of the goal of using RDF metadata with URIs was to enable seamless access to different metadata catalogues without having to deal with specific standards. In particular OGC metadata (19115/39 used by INSPIRE), EML metadata (GBIF), Bibliographic references, Dublin Core metadata have been transformed to comply with a common set of metadata elements (cf section

3.1). Moreover these metadata are all anotated (and thus linked) with URIs of domain entities (cf section 3.2). This approach enables to query resources related to domain entities (e.g. yellowfin tuna) without having to restrict the search to specific standards, languages or terms. The search engine suggests all the results related to this entity and cluster the results according to their types Results can be related entities (e.g. preys of yellowfin tuna) or information resources like pictures, articles, databases, people... (see online application).

Fig. 5. Search engine for the Ecoscope knowledge base

More than inventorying existing information resources, our prior goal was to summarize available knowledge by themes / domain entities. This has been done by setting up fact sheets.

5.2 Fact sheets

The main purpose of our knowledge base was to feed the content of a Web portal by providing the knowledge about entities of interest for our laboratory (ecosystems, species, fishing vessels...). The main goal was the creation of fact sheets about these entities. To achieve this goal, A SPARQL query harvest all the triples related to a given entities and Jena objects are used by Struts2 to build some HTML views. Figure 6 gives an example of Web page for yellowfin tuna.

The fact sheets cluster related entities by type of relationships (*is predator of*, *is prey of*, *is exploited by*) and cluster related information resources by data types (pictures, spatial data and related processes / indicators).

Other visualization interfaces are available to present RDF triples as taxonomy or network.

Fig. 6. Fact sheet about Yellowfin tuna

5.3 Visualization of a food web

This use case illustrates what can be achieved with previous ontologies when applied to management of biodiversity data. The example of a food web is very relevant as it shows relationships between entities (species) that are either predators, preys or both. In Figure 7, we filtered RDF triples of the knowledge base to represent the food web related to tropical tuna trophic data (using prefuse API [8] for visualization of data). This application is interfaced with relational databases in order to enable users to spatially query the food web.

Fig. 7. Representation of a food web from RDF triples

The Figure 6 gives an example of Web page which is available online with other visualization application (e.g. Taxonomy).

5.4 Matching service

Another use case consists in using the Ecoscope knowledge base for the mapping of codification systems (code lists). For example, in EAF domain, species are often identified either by FAO codes in fisheries datasets or, most of the time, by Worms codification systems in ecological observations. This is an issue when researchers need to run some processes which require cross analyses between fisheries and ecological datasets. Indeed, in this case, there is a need to enable mapping between codes to merge datasets. A first application has been developed to enable such mapping at data export. We aim to deliver similar services in a generic way (in a programmatic way or through GUIs) to enable mapping between code lists of different schemas.

6 Outlooks

For now, RDF triples to link our data with entities and agents of the domain have been created in various ways. To make the knowledge base sustainable in the long term, we can't afford to feed it with ad hoc approaches. There is a need to harvest information from dedicated endpoints to facilitate updates by a workflow:

- databases and netCDF files will be turn in RDF through a single data server,
- pictures that are collected by on-board observers or scientists,
- articles, reports... published by researchers will be exported in RDF from Zotero Server,
- processes to run data analysis with different programming languages (R, IDL, Matlab...) will be described in RDF from OGC WPS metadata.

Another important improvement is related to the introduction of logical rules to infer some knowledge. A simple example consists in infering *competition* relationship between species from *predation* relationship. Indeed, two species are competing when they are predators of the same species (preys). This first use case is going to be implemented in the framework of iMarine FP7 program.

7 Conclusion

Our first application has demonstrated the interest of using ontologies and knowledge bases to satisfy different needs of researchers in our marine ecology laboratory: data discovery and retrieval, knowledge management and visualization (fact sheets, food web...). Such an application is worth but our current "ad hoc" approach still requires a lot of work to fill and update the database. To fix this issue, we aim to enable a RDF export from the Web portal which is used to access raw data (relationnal databases and netCDF files). We aim now to set up a workflow to facilitate RDF generation directly from the relational databases where ecological observations are stored and used by researchers to run scientific analysis. As a second step, we want ecological observations (raw data) to be

made available as well in RDF and to be linked with existing RDF triples (summarizing underlying knowledge: for example a RDF triple stating a predation relationship between two species should be related to hundreds of observations / facts proving it). The next goal consists in enabling RDF export from our main data sources, as already done with other standards (OGC, TDWG, SDMX).

References

1. Tim Berners-Lee. Linked data. July 2006.
2. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
3. Dan Brickley and Libby Miller. Foaf vocabulary specification. Namespace document, January 2010.
4. DCMI. Dublin core metadata initiative. *http://dublincore.org/*.
5. FAO. The ecosystem approach to fisheries. *FAO Guidelines for Responsible Fisheries*, (4):112, 2003.
6. FAO. Fisheries linked open data. *http://www.fao.org/figis/flod/*, 2011.
7. Frédérick Giasson and Bruce D'Arcus. Bibliographic ontology. Technical report.
8. Prefuse information visualization toolkit. http://prefuse.org/.
9. Joshua Madin, Shawn Bowers, Mark Schildhauer, and Matthew Jones. Advancing ecological research with ontologies. *Trends in Ecology & Evolution*, 23(3):159–168, March 2008.
10. Joshua Madin, Shawn Bowers, Mark Schildhauer, Sergeui Krivov, Deana Pennington, and Ferdinando Villa. An ontology for describing and synthesizing ecological observation data. *ECOLOGICAL INFORMATICS*, 2(3, Sp. Iss. SI):279–296, October 2007.
11. Alistair Miles and José R. Pérez-Agüera. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83, 2007.
12. Trina S. Myers, Ian Atkinson, and Ron Johnstone. Supporting coral reef ecosystems research through modelling re-usable ontologies. In *Knowledge Representation Ontology Workshop*, Proceedings of Conferences in Research and Practice in Information Technology (CRPIT), September 2008.
13. OGC. Open geospatial consortium, http://www.opengeospatial.org/.
14. Cynthia S Parr and Michael P Cummings. Data sharing in ecology and evolution. *Trends in Ecology & Evolution (Personal Edition)*, 20(7):362–363, July 2005. PMID: 16701396.
15. O. J. Reichman, Matthew B. Jones, and Mark P. Schildhauer. Challenges and Opportunities of Open Data in Ecology. *Science*, 331(6018):703–705, February 2011.
16. Leo Sauermann, Richard Cyganiak, and Max Völkel. Cool uris for the semantic web. Technical Memo TM-07-01, DFKI GmbH, Deutsches Forschungszentrum f ür Künstliche Intelligenz GmbH, February 2007. Written by 29.11.2006.
17. SDMX. Statistical data and metadata exchange. *http://sdmx.org/*, 2011.
18. TDWG. Taxonomic database working groug. *http://www.tdwg.org/*.
19. Jo Walsh and Pedro Goncalves. Dclite4g vocabulary. *http://dclite4g.xmlns.com/*.
20. John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: an evolving community-developed biodiversity data standard. *PLoS One*, 7(1):e29715, 2012.
21. Richard J. Williams, Nea D. Martinez, and Jennifer Goldbeck. Ontologies for ecoinformatics. *Journal of Web Semantics*, 4:237–242, 2006.