Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

RITMARE: Semantics-aware Harmonisation of Data in Italian Marine Research

Cristiano Fugazza[a], Anna Basoni[a], Stefano Menegon[b], Alessandro Oggioni[a], Fabio Pavesi[a], Monica Pepe[a], Alessandro Sarretta[c], Paola Carrara[a]

[a]*Istituto per il Rilevamento Elettromagnetico dell'Ambiente – CNR, v. Bassini 15, 20133 Milan, Italy*
[b]*Istituto di Scienze Marine – CNR, Arsenale - Tesa 104, Castello 2737/F, 30122 Venice, Italy*
[c]*Istituto di Scienze Marine – CNR, v. Gobetti 101, 40129 Bologna, Italy*

Abstract

RITMARE is a Flagship Project by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) and coordinated by the National Research Council (CNR). It aims at the interdisciplinary integration of national marine research. In pursuing a Linked Open Data (LOD) vocation, the RITMARE sub-project 7 is building the necessary domain-related data structures by leveraging existing RDF-based schemata and sources. These data structures are grounding semantics-aware profiling of end users, data providers, and resources. The goal is designing a flexible infrastructure that adapts to the audience's specificities.

Peer-review under responsibility of euroCRIS.

Keywords: Spatial Data Infrastructures; Semantic Web; Marine Research; RDF

1. Introduction

Spatial Data Infrastructures (SDIs) are applications for the collection and provisioning of geospatial data, metadata, networked services, and technologies. Differently from generalised search, where indexing and retrieval is driven by crawling (hyper)textual data on the Web, this category of data relies on specific metadata formats to effectively "discover" resources. In this context, recourse to semantics is typically intended to address heterogeneity in resource description, to achieve multilingualism in resource discovery functionalities, and in general to compensate for the lack of the efficient search paradigms of generalized information retrieval.

Instead, it is our opinion that, in order to fully exploit the potential of the fine-grained descriptions enabled by ontologies, thesauri, and RDF-based data structures in general[10], it is necessary to adopt an all-round approach encompassing all the aspects of SDIs. In this paper, we report on the activities that are carried out for the

Peer-review under responsibility of euroCRIS.

establishment of the RITMARE infrastructure with regard to semantics. In a nutshell, we employ heterogeneous schemata to integrate in a coherent knowledge base the information that is required for the enactment of the infrastructure.

The paper is organised as follows. Section 2 describes RITMARE, particularly with regard to the requirement analysis that set the groundwork for the development of an interoperable infrastructure. Section 3 describes the RDF data structures that constitute the project's primary metadata store, also providing examples of how these can enable advanced discovery capabilities. Finally, Section 4 draws conclusions and provides an outlook on future developments.

2. The RITMARE Flagship Project

The RITMARE[13] (la Ricerca ITaliana per il MARE - Italian research for the sea) Flagship Project is one of the National Research Programmes funded by the Italian MIUR. The project addresses the whole sector of marine research, involving a number of public research bodies and inter-university consortia as well as private companies. RITMARE has been structured around the following three objectives:

1. Supporting integrated policies for the safeguard of the environment (the health of the sea);
2. enabling sustainable use of resources (the sea as a system of production);
3. implementing a strategy of prevention and mitigation of natural impacts (the sea as a risk factor).

The project is organised into seven sub-projects and, among them, sub-project 7 aims at building an interoperable infrastructure for the observation network and marine data. The infrastructure shall be capable of interconnecting the whole community of researchers involved in RITMARE. It will allow coordinating and sharing of data, processes, and information produced by the other sub-projects. A main pre-requisite is not to hamper existing practices and enabling technologies adopted by the individual scientific communities. In fact, the great variety of actors reflects in the coexistence, within the project, of different data formats, practices, approaches, and requirements: RITMARE involves thousands of researchers with different scientific backgrounds, technological skills, and objectives. They can be roughly grouped in the following disciplinary communities:

- Physical oceanography
- Chemical oceanography
- Geology
- Geophysics
- Coastal systems
- Ecology
- Fishery and aquaculture
- Biomolecular science
- Human impacts
- Climatology
- Biogeochemistry

All of them use different types of data, described by heterogeneous metadata, managed by own workflows; usage of heterogeneous vocabularies is not infrequent.

2.1. Requirement analysis and baseline survey

The RITMARE information infrastructure tackles the challenge of building a network of interconnections and tools to help data flows, to ease information and service management, to meet the needs of the scientific communities involved, and to support their growth. In order to achieve this, within its first year, sub-project 7 performed multiple analysis activities; among these, the collection of requirements from the project's researchers with respect to data and data process infrastructure. In particular, this activity aimed at defining:

- The data to be used, managed, and produced;
- their characteristics;
- the functionalities and tools that are required to work on data;
- the workflows to be supported.

The requirement collection was carried out by means of interviews to a representative sample of 30 selected RITMARE researchers, who expressed a grand total of 104 requirements. The output of interviews are stored in a publicly accessible archive[12]. The analysis of requirements revealed that the project's scientific community is pretty unanimous in recommending that:

- The infrastructure should manage a wide spectrum of data typologies (such as digital terrain models, bathymetries, remote sensing imagery, bio-molecular data and processes, environmental observations, etc.) produced both within and outside RITMARE;
- the data made available through the infrastructure should be shared and managed by researchers in order to, for example, feed models, support decisions, and plan research activities.

Requirements have been clustered in 6 macro-requirements that have been examined and evaluated by an expert panel in order to establish their relative importance and priority. At the same time, another group in sub-project 7 performed an analysis of the state-of-the-art solutions adopted by the marine researchers for the purpose of managing their data and workflows. The information that has been collected has been summarized in sheets describing more than 60 initiatives in the field of data infrastructures for marine research. The descriptions include the following fields:

- Identification information and main functionalities addressed by the initiative/project;
- a description of the data and information made available, used, and managed;
- the maturity level of the initiative/project with respect to the objectives of RITMARE;
- the hardware and software deployed in the surveyed infrastructures.

3. Towards a semantics-aware architecture

The landscape of requirements and existing solutions pinpointed by the preliminary analysis proved to be extremely heterogeneous and challenging w.r.t. state-of-the-art geoportal approaches. In fact, traditional online access, with uniform tools for the whole users' arena, often hampers friendly collaboration among the researchers and fair support of their interaction needs. This is the rationale for relying on semantics in order to ease management of resources in RITMARE and achieve a flexible overall solution. In fact, the infrastructure is meant to overcome interoperability, interdisciplinary, and IT unfriendliness issues via semantic tools that are tailored on the habits, skills, and needs of the researchers that are involved in the project.

Among the guidelines that have been set for the engineering of the RITMARE infrastructure is a sound foundation of data structures on Semantic Web standards and practices. The rationale for this is aligning the project with Italian legislation with regard to Liked Open Data (LOD)[1] and bridging the gap between the management of spatial information (as mandated by the INSPIRE Directive[2] and its Italian counterpart[3]) and the variety of Open Data initiatives that flourished since the initial formulation of the Directive in 2007.

In fact, because of this gap, the implementation of INSPIRE fell short of taking advantage of the large number of semantics-aware data structures that coalesced in the following years into the multi-tenanted knowledge base that is typically referred to as the "LOD cloud"[11]. The first step for achieving this was expressing the context information of required by the project as RDF; this was achieved by leveraging widely-acknowledged schemata that were integrated into a consistent knowledge base. These components are portrayed in Figure 1 and consist of:

- The categorisation of researchers, research institutes, and the project's internal, multi-level organisation as FOAF (Friend Of A Friend) data structures (around 1.800 entities)[7];

- an exhaustive collection of SKOS (Simple Knowledge Organization System) thesauri[9] featuring Earth Observation parameters and measuring units[4], the research domains associated with the project, and the code lists used for assisted metadata editing (around 30.000 entities);
- a collection of features related to Italy or the Mediterranean sea taken from the GeoNames ontology and knowledge base (around 40.000 entities)[8];
- finally, the essential component of the architecture is the metadata of spatial resources in a format that is compatible with the RDF-based representation of resources in the EU Open Data Portal[5] (based on the DCAT data format by W3C[6]). The prescribed ISO-based representation can be generated on-demand.

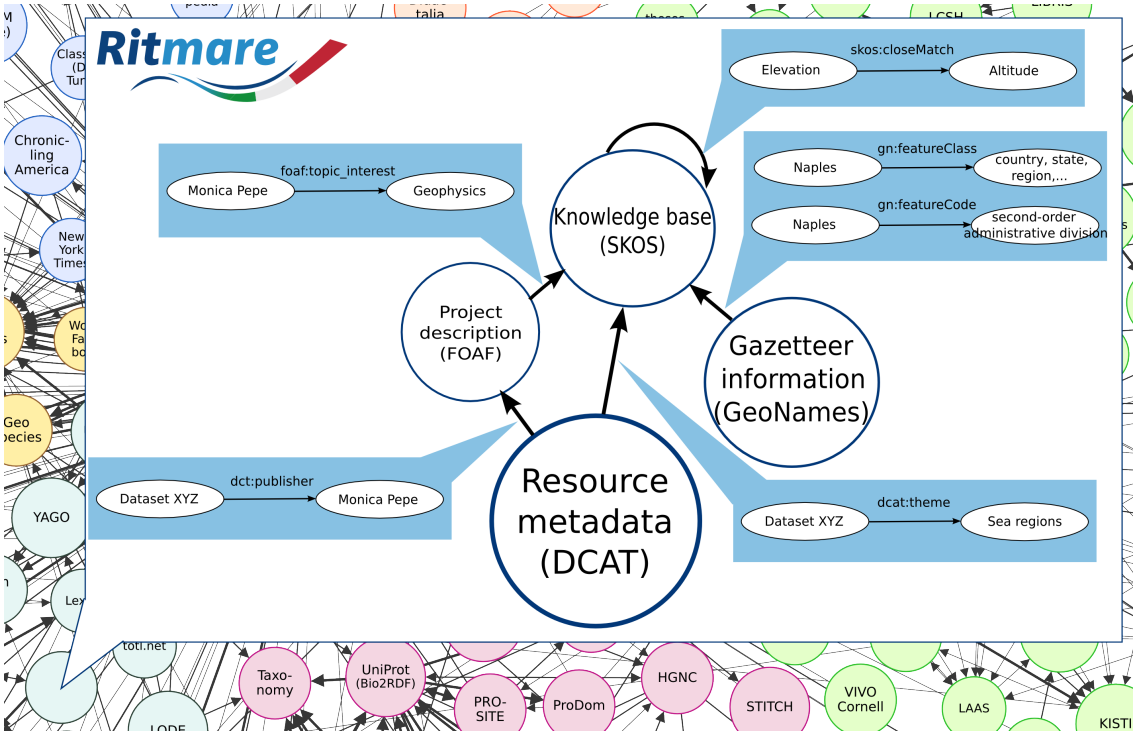

Fig. 1. RDF knowledge base underpinning the RITMARE infrastructure.

The challenge is to implement INSPIRE-compliant catalogues by the resource providers involved in the project and, at the same time, enable more fine-grained resource discovery criteria by the central repository of metadata. These data structures (which will be refined and extended throughout the project's lifespan) can be seamlessly integrated with each other by means of properties defined by the underlying schemata. More specifically:

- FOAF entities are related to SKOS concepts through FOAF property *topic_interest*: As an example, Figure 1 shows the FOAF entity corresponding to researcher “Monica Pepe” linked to the SKOS concept corresponding to “Geophysics”;
- heterogeneous SKOS thesauri are mapped one onto the other by means of the semantic properties defined by SKOS: in Figure 1, the SKOS concepts corresponding to “Elevation” and “Altitude” are related to each other by semantic property *closeMatch*;
- gazetteer data structures are already referring to controlled vocabularies, also included in the knowledge base, for categorizing features. As an example, Figure 1 shows two SKOS concepts (described by literals “country, state, region, ...” and “second-order administrative division”) that are referred to by the feature corresponding to

"Naples" as values for, respectively, properties *featureClass* and *featureCode* defined by the GeoNames ontology;

- DCAT metadata items reference entities from the above data structures and from the gazetteer in order to specify properties such as resource creator, topic, geographic extent, etc.: In the example in Figure 1, the dataset "Dataset XYZ" has been related to the INSPIRE Theme "Sea regions" and to the publisher "Monica Pepe", respectively, by properties *theme* and *publisher* from the DCAT vocabulary at metadata creation-time.

These categories of semantic links are established in order to improve the user experience in the specific phase of discovery of spatial resources but also to increase usability of the overall access interface to the infrastructure. More specifically, the first category allows for advanced metadata creation functionalities, enable profiled discovery of resources, and tailor the selection of components to be included in the infrastructure's main interface on the user's specificities. Instead, the second category of semantic link is a fundamental enabling factor for query expansion functionalities that, given the inter-disciplinary vocation of RITMARE, may be an important component for the usability of the infrastructure. Multilingualism can also be achieved by selecting terminologies that are made available in multiple languages. Finally, linking datasets to the toponyms in GeoNames allows to express query patterns like "100km west of Naples" and reduce recourse to maps, bounding boxes, and other widgets that typically distinguish geographic data discovery from generalized search.

4. Conclusions

This paper described the RDF-based data structures that are to be employed in the creation of a centralised repository of metadata from the heterogeneous data sources in the RITMARE research network. These allow to implement a user-profiled interface to the infrastructure and enable semantics-aware discovery capabilities that take into account the semantic relationships among the entities that are referred to in resource descriptors.

We believe that, by relating metadata records to the entities populating context information, as described in Section 3, it is possible to better support the end-user throughout the whole life-cycle of spatial resources, from metadata creation to discovery, and contribute to the geospatial LOD cloud.

The RITMARE sub-project 7 is now engaged in an effort aimed at enabling researchers to create standard-based data repositories supported by an adequate set of metadata.

Acknowledgements

The activities described in this paper have been funded by the Italian Flagship Project RITMARE.

References

1. Commissione di Coordinamento SPC. (2013). Linee guida per l'interoperabilità semantica attraverso i Linked Open Data. Online <http://archivio.digitpa.gov.it/notizie/linee-guida-open-data-interoperabili>
2. European Commission, "Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)" Directive 2007/2/EC, Official J. European Union, vol. 50, no. L 108, 2007, pp. 1–14.
3. Presidenza del Consiglio dei Ministri – Agenzia per l'Italia Digitale. Regole tecniche Repertorio nazionale dati territoriali – DM 10 novembre 2011.
4. SeaDataNet - BODC webservices. Online <http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx>
5. European Commission, "DCAT application profile for data portals in Europe". Online <https://joinup.ec.europa.eu/asset/dcat_application_profile>
6. W3C, "Data Catalog Vocabulary (DCAT)". Online <http://www.w3.org/TR/vocab-dcat/>
7. FOAF Project.Online <http://www.foaf-project.org/>
8. GeoNames geographical database. Online <http://www.geonames.org/>
9. Simple Knowledge Organization System (SKOS) website. Online <http://www.w3.org/2004/02/skos/>
10. Resource Description Framework (RDF). Online <http://www.w3.org/RDF/>
11. LOD Cloud. Online <http://linkeddata.org/>
12. RITMARE SP7 requirement anlysis. Online <http://sp7.irea.cnr.it/wp1/az1/questresults/interviste.php>
13. RITMARE Flagship Project. Online <http://ritmare.it>