Streamlining geospatial metadata in the Semantic Web

Cristiano Fugazza[1], Monica Pepe[1], Alessandro Oggioni[1], Paolo Tagliolato[1 2 3], and Paola Carrara[1]

[1] Institute for Electromagnetic Sensing of the Environment, National Research Council (IREA-CNR), via Corti 12, 20133 Milan, Italy
[2] Institute of Marine Science, National Research Council (ISMAR-CNR), Tesa 104 Arsenale, Castello 2737/F, 30122 Venice, Italy
[3] LifeWatch Italy

E-mail: {fugazza.c, pepe.m, oggioni.a, tagliolato.p, carrara.p}@irea.cnr.it

Abstract. In the geospatial realm, data annotation and discovery rely on a number of ad-hoc formats and protocols. These have been created to enable domain-specific use cases generalized search is not feasible for. Metadata are at the heart of the discovery process and nevertheless they are often neglected or encoded in formats that either are not aimed at efficient retrieval of resources or are plainly outdated. Particularly, the quantum leap represented by the Linked Open Data (LOD) movement did not induce so far a consistent, interlinked baseline in the geospatial domain. In a nutshell, datasets, scientific literature related to them, and ultimately the researchers behind these products are only loosely connected; the corresponding metadata intelligible only to humans, duplicated on different systems, seldom consistently.

Instead, our workflow for metadata management envisages i) editing via customizable web-based forms, ii) encoding of records in any XML application profile, iii) translation into RDF (involving the semantic lift of metadata records), and finally iv) storage of the metadata as RDF and back-translation into the original XML format with added semantics-aware features. Phase iii) hinges on relating resource metadata to RDF data structures that represent keywords from code lists and controlled vocabularies, toponyms, researchers, institutes, and virtually any description one can retrieve (or directly publish) in the LOD Cloud. In the context of a distributed Spatial Data Infrastructure (SDI) built on free and open-source software, we detail phases iii) and iv) of our workflow for the semantics-aware management of geospatial metadata.

1. Introduction

The implementation of the INSPIRE (INfrastructure for SPatial InfoRmation in Europe) Directive [1, 2, 3] provided the geospatial community with recipes for interoperability that have sound foundation on standards. However, the landscape of data management has inevitably changed since the formulation of the Directive and its early, prototype implementations. Nowadays, the once-state-of-the-art groundwork set by the reference ISO standards for geographic information (such as [4]) has become inadequate, primarily because of the change in perspective from data representation (that is, the quest for appropriate encoding mechanisms for data and metadata in a specific domain) to data access and processing. As an example,

terms like Open Data [5], Linked Data [6], RDF[1], and SPARQL[2] are increasingly associated with data interoperability and open-ness.

The engineering of the data sharing infrastructure in RITMARE[3] (a Flagship Project by the Italian *Ministero dell'Istruzione, dell'Università e della Ricerca*) requires to bring all state-of-the-art contributions to marine research under the same umbrella. In tackling this heterogeneity, we decided to adopt best practices that, besides fostering interoperability [7], could also ease migration to the aforementioned novel paradigms for data representation and access: Specifically, we implemented a workflow for creation of semantics-aware metadata in order to integrate the resource descriptions from the individual data sources into a homogeneous catalog that supports single-entry-point discovery. The output of this research thread is made available as FOSS on GitHub[4].

Our purpose is twofold: On the one hand, we need to harmonize metadata according to the baseline set by the INSPIRE Directive and the practices set by Sensor Web Enablement (SWE)[5] [8, 9, 10]. On the other, we want to ground metadata creation on controlled vocabularies and semantics-aware context information in order to bridge the gap between legislative compliance, Linked Open Data fruition practices, and the Italian guidelines on the subject [11]. This paper addresses the second thread in the development of the RITMARE infrastructure as regards metadata, that is, the structuring of semantics-aware metadata.

The remaining of the paper is organized as follows. Section 2 describes the limitations of state-of-the-art metadata representation and presents a simple use case. This Section also presents some related work that is relevant to the paper. Section 3 introduces the varied RDF-based schemata that express the project's context information (such as users, terminologies, toponyms, etc.) and that is subsequently employed to assist metadata creation. It also provides some details on the templating language driving the creation of metadata editing interfaces harnessing this information. In this Section, the improvement over the use case presented in Section 2 stemming from our workflow is made apparent. Finally, Section 4 draws conclusions and outlines future work in the establishment of the RITMARE infrastructure as regards metadata.

2. Context

Geospatial metadata is largely based on free-text descriptions. In developing our tools for assisted metadata creation, we explored a broad range of formats, from Unidata's NcML[6] that is used for annotating NetCDF resources made available by THREDDS servers to the SensorML metadata employed in SWE. This choice of example formats is functional to marking the two ends of such spectrum. In fact, NcML is relying on the Climate and Forecast (CF) conventions[7] to enable some degree of interoperability among catalogs and resources but, by the end of the day, code values are inserted in the metadata as free text. On the contrary, SensorML provides a number of XML attributes for complementing free-text property values with unambiguous identifiers, such as Uniform Resource Identifiers (URIs)[8]. The INSPIRE schema we are referring to in this paper is an in-between example of metadata format inasmuch it defines (or inherits from ISO 19115/19119) a number of codelists to choose property values from for inclusion (again, as free text) in descriptions.

[1] Resource Description Framework (RDF): `http://www.w3.org/RDF/`
[2] SPARQL Query Language for RDF: `http://www.w3.org/TR/rdf-sparql-query/`
[3] RITMARE (Ricerca ITaliana per il MARE - Italian research for the sea): `http://www.ritmare.it/`
[4] SP7 Interoperable Data Infrastructure for RITMARE: `https://github.com/SP7-Ritmare`
[5] Sensor Web Enablement (SWE): `http://www.opengeospatial.org/projects/groups/sensorwebdwg`
[6] The NetCDF Markup Language (NcML): `http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/ncml/`
[7] NetCDF CF Metadata Conventions: `http://cfconventions.org/`
[8] Uniform Resource Identifier (URI): Generic Syntax: `https://tools.ietf.org/html/rfc3986`

Listing 1. Metadata fragment defining John Doe as creator of resource "Dataset ABC"

```
01 <gmd:pointOfContact>
02   <gmd:CI_ResponsibleParty>
03     <gmd:organisationName>
04       <gco:CharacterString>ACME Research</gco:CharacterString>
05     </gmd:organisationName>
06     <gmd:role>
07       <gmd:CI_RoleCode codeList="..." codeListValue="author"/>
08     </gmd:role>
09     <gmd:contactInfo>
10       <gmd:CI_Contact>
11         <gmd:address>
12           <gmd:CI_Address>
13             <gmd:electronicMailAddress>
14               <gco:CharacterString>
15                 john.doe@acmeresearch.org
16               </gco:CharacterString>
17             </gmd:electronicMailAddress>
18           </gmd:CI_Address>
19         </gmd:address>
20       </gmd:CI_Contact>
21     </gmd:contactInfo>
22   </gmd:CI_ResponsibleParty>
23 </gmd:pointOfContact>
```

Moreover, a number of property values can not be constrained by code lists (e.g., the e-mail address of a contact point) and nevertheless these can be related to authoritative, unambiguously identified data structures (e.g., for contact points, this could be the URI of the ORCID profile[9] of a researcher, retrieved as RDF through content negotiation). Looking up these data structures to retrieve the up-to-date values for a given property represents a viable solution to metadata inconsistency, which is a major issue the geospatial community, among the others, will be required to address at some point. In order to clarify the problem we are addressing, consider this simple use case: "John Doe" is a researcher by "ACME Research", an institute managing geospatial resources through an INSPIRE-compliant catalog service. Specifically, he is the creator of the dataset shown in Figure 1, "Dataset ABC", made available by the institute's SDI. The metadata fragment conveying this information is shown in Listing 1.

At some point, John Doe changes employer and, as a consequence of this, the metadata record for resource "Dataset ABC" is made inconsistent: The name of the institute (line 4) has to be changed, as well as the e-mail address of the individual (line 15). Such update has to be carried out manually, a process that can be error-prone or simply overlooked. Instead, we need an approach enabling an unambiguous identification and description of both the individual and his new employer, so that the aforementioned property values can be retrieved when the metadata record is actually requested. Recourse to RDF-based context information provides this feature and also allows for supporting field editing and context-based validation. Specifically, the solution we are investigating entails a metadata editing tool that produces descriptions according to generic XML-based metadata profiles by relying on a template language, a metalanguage. Please refer to [12] for further information on this tool. The latter allows for a threefold exploitation of RDF data structures; in particular:

[9] ORCID - Connecting Research and Researchers: `http://orcid.org/`

- Generic RDF data sources can be plugged in, provided they are made available as a SPARQL endpoint. Property values for metadata items can then be filled in on the basis of query results.
- Besides the XML profile the editing tool is configured to produce, a semantics-aware version of the metadata record can be created.
- The metadata in the original XML profile can be reconstructed on demand, thus relying on up-to-date information for reducing as possible inconsistency issues.

Relating property values to unique identifiers improves overall interoperability of infrastructures because, on the one hand, it elicits processing by automated agents and, on the other, it allows catalogs to overcome discovery issues related to multilingualism and semantic heterogeneity. Still our intent goes farther because, by adopting RDF as the native metadata storage format and making extensive use of URIs as property values, we foster management of geospatial metadata in a distributed, multi-tenanted fashion. In fact, the Semantic Web is a distributed corpus of information whose RDF-based resources (the entities identified by URIs) can often be directly accessed as Linked Data, thus it makes little sense sticking to the notion of metadata as data structures that are single-handedly maintained by the institution providing a given resource.

2.1. Related work

As reuse of existing schemata is a mantra in contributing to the Semantic Web, our primary reference when working out the translation of ISO-based geospatial metadata into RDF has been the OWL representation of the ISO/TC 211 UML Model for geographic information[10] by the Australian *Commonwealth Scientific and Industrial Research Organization* (CSIRO). Another source of direction is the Data Catalog Vocabulary (DCAT)[11] by W3C, the data schema underlying the existing Open Data portals based on CKAN[12]. Interestingly, the DCAT application profile for data portals in Europe is featuring an extension for representing INSPIRE metadata[13] that has recently reached maturity for public review. Once the baseline is set with the DCAT schema (to ease integration with portals based on CKAN), whether to fill in the missing metadata elements with either of the aforementioned contributions to the state of the art was just a matter of taste. For the time being, we are sticking to the former.

3. Semantic lift of resource metadata

Figure 1 depicts the individual components of the RDF data structures that are being employed in developing the RITMARE infrastructure (ellipses containing mnemonic labels that stand for longer, less intelligible URIs). In order to exploit these data structures in a seamless way, it is important to provide the appropriate relations between entities from distinct schemata, as exemplified by the sample relations that connect the four bubbles.

Project Description Researchers and institutions involved in the project are modeled as Friend Of A Friend (FOAF) data structures[14]. As an example, Figure 1 shows the FOAF entity corresponding to researcher "John Doe" linked to the SKOS concept corresponding to "Geophysics" by property *topic_interest* from the FOAF vocabulary.

Knowledge Base The second component is constituted by a collection of terminologies in the Simple Knowledge Organization System (SKOS) format that are employed in the

[10] ISO/TC 211 geographic information as OWL: `http://def.seegrid.csiro.au/static/isotc211/`
[11] Data Catalog Vocabulary (DCAT): `http://www.w3.org/TR/vocab-dcat/`
[12] CKAN - The open source data portal software: `http://ckan.org/`
[13] GeoDCAT-AP working drafts: `https://joinup.ec.europa.eu/node/139283/`
[14] FOAF Vocabulary Specification 0.99: `http://xmlns.com/foaf/spec/`

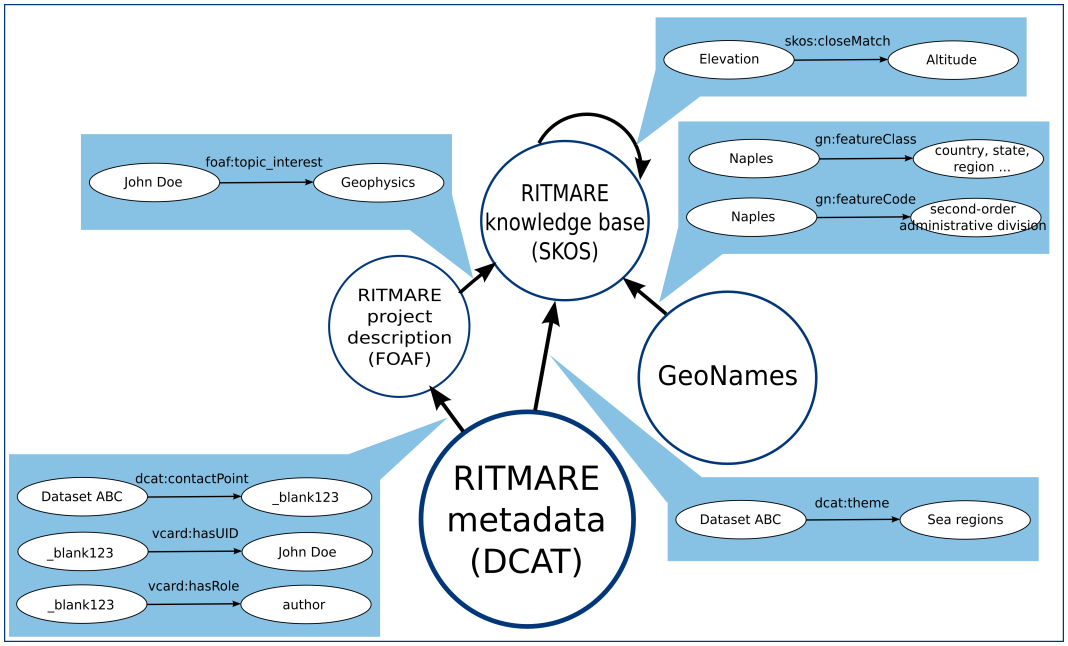

Figure 1. Context information in the RITMARE infrastructure

creation of INSPIRE-compliant metadata. Figure 1 portrays the mapping between concepts "Elevation" and "Altitude" expressed by semantic property *closeMatch* from the SKOS ontology.

Gazetteer Information Another adopted data structure is a fine-grained representation of the geographic features that may be referred to in user queries. Gazetteer data structures are also referring to controlled vocabularies, such as the categorization of GeoNames'[15] toponym "Naples" shown in Figure 1.

Resource Metadata Finally, the largest proportion of the RDF data employed by the RITMARE infrastracture is constituted by resource metadata. In the example in Figure 1, the resource "Dataset ABC" has been related to the INSPIRE Theme "Sea regions" by property *theme* from the DCAT vocabulary. The dataset is also related to the creator of the resource by introducing the blank node "_blank123" to reference entity "John Doe" and a SKOS concept standing for the role defined by ISO for "author".

The data structures that have been described above (and any other SPARQL-compliant data source) can be used for assisting the editing of metadata. This is the primary means for relating metadata items to resources in the Semantic Web. Listing 2 shows a fragment of the template for INSPIRE-compliant metadata highlighting the logic underlying the three usages enumerated in Section 2.

Specifically, lines 2-15 define datasource "person" (referred to later in line 24) that allows the interface to autocomplete the metadata field corresponging to the e-mail address of the point of contact (lines 20-34 of Listing 2): The SPARQL query that is defined for the datasource contains parameter "$search_param" that is replaced by the text pattern entered by the user in the corresponding form field. This is one of the many ways our template language may assist the user in providing the metadata: As an example, the field in lines 35-48 is autocompleted on the basis of the user that has been selected in the step we just introduced (the mechanics of this omitted in Listing 2 for brevity); metadata fields can also be autocompleted on the basis of

[15] GeoNames: `http://www.geonames.org/`

Listing 2. Code fragment from the INSPIRE template

```
<template>
  <datasource>
    <id>person</id>
    <query>
    <![CDATA[
      SELECT ?c ?l
      WHERE {
        ?c rdf:type foaf:Person .
        ?c vcard:email ?l .
        FILTER( REGEX( STR(?l), "$search_param", "i") )
      }
      ORDER BY ASC(?l)
    ]]>
    </query>
  </datasource>
  <element>
    <id>resp</id>
    <label xml:lang="en">Responsible party</label>
    <produces>
      <item>
        <hasIndex>1</hasIndex>
        <label xml:lang="en">Email</label>
        <hasPath>/.../gmd:electronicMailAddress/...</hasPath>
        <datasource>person</datasource>
        <RDFout>
        <![CDATA[
          <$id_md_1_uri> dcat:contactPoint [
            vcard:hasUID <$resp_1_uri> ;
            vcard:hasRole <$resp_3_uri>
          ] .
        ]]>
        </RDFout>
        <RDFin>...</RDFin>
      </item>
      <item>
        <hasIndex>2</hasIndex>
        <label xml:lang="en">Institute</label>
        <hasPath>/.../gmd:organisationName/...</hasPath>
        <RDFin>
        <![CDATA[
          SELECT ?i
          WHERE {
            <$resp_1_uri> vcard:org ?o .
            ?o foaf:name ?i .
          }
        ]]>
        </RDFin>
      </item>
    </produces>
  </element>
</template>
```

HTTP parameters in the call to the editing interface, can be generated on demand, can duplicate the content of another field, and even use generic XPath functions in order to mix-and-match values in other form fields.

Once metadata is posted to the server-side component of the tool, the prescribed XML output is generated by creating XML elements and attributes following XPath expressions (such as those in lines 23 and 38). Additionally, element *RDFout* defined by the template language allows to specify which RDF triples shall be produced and inserted in the triple store that is used for metadata storage. As an example, lines 25-32 contain the *blueprint* for creating the point of contact in the bottom-right part of Figure 1 as RDF in the Turtle[16] syntax. Parameters in the form "$<parameter_name>" follow the pattern "$<element_id>_<item_index>[_uri]": As an example, parameter "$resp_1_uri" looks up the first item in element "resp" (the e-mail address of the point of contact) and retrieves the corresponding URI (the unique identifier for the user that have been selected). Note that no element *RDFout* is defined for the second item in the template: The rationale for this is that no RDF data shall be generated to express the institute name, because this value can be retrieved on demand (e.g., when the metadata of the resource is looked up). This is the major difference between our approach and any of the possible one-to-one translations of INSPIRE metadata into RDF.

In fact, element *RDFin* in the template definition for item #2 (lines 39-47) contains the SPARQL query that shall be executed to retrieve the up-to-date value of the metadata item. The query can be matched against the data source that is defined for the metadata item in hand (omitted in Listing 2 for brevity). Albeit the query is supposed to return no hits (because the record associated with John Doe has been removed fro the data structures in Figure 1), the SPARQL query contains all necessary information to try a number of Linked Data lookups (in this case, one for the URI expressed by "resp_1_uri" and one for the URI that is returned by the first lookup). The query is straightforward and assumes that the target data source expresses the intended pieces of information using the *vcard* and *foaf* schemata; however, it could be made more general by allowing for multiple schemata, alternative values, etc. Once the metadata record corresponding to a resource is requested, the template is parsed and the distinct metadata items are filled in with the values that are retrieved, either from the data source that is defined or from the Semantic Web at large. Then the prescribed XML format is generated.

Referring to the use case presented in Section 2, John Doe changing research institute, our RDF representation of metadata involves the information in the bottom-right part of Figure 1, generated by the *RDFout* block in lines 25-32 of Listing 2. Assume that the URI points to John Doe's FOAF profile, which has been updated by the researcher in order to point to the FOAF profile of the new employer. Now the metadata for resource "Dataset ABC" can be produced on the fly: It will be analogous to that in Listing 1, except that the values of the two items that are inconsistent in the original, hard-coded metadata are automatically updated to the new values.

4. Conclusions and future work

In this paper, with reference to the workflow for metadata management in project RITMARE, we outlined the solutions adopted to ground metadata creation on the Semantic Web. Besides the prescribed XML representation that can be processed by CSW-compliant and SWE applications, we feature an RDF representation that allows for enriching metadata with semantic information that can be used at a later stage in order to provide increased features. We are structuring the discovery and presentation functionalities taking advantage of this additional knowledge; these are going to be exposed by the geoportal for project RITMARE currently under development.

Let aside providing extended discovery functionalities, such as those described in [13, 14],

[16] Terse RDF Triple Language (Turtle): `http://www.w3.org/TeamSubmission/turtle/`

our methodology fosters distributed management of metadata. This practice can impact on the capabilities of geoportals as regards maintaining long-term consistency of metadata records, as exemplified by the simple use case presented in this paper. We stress the importance of this aspect because of the fluidity of research environments that can not be easily coped with by state-of-the-art geoportals. In fact, current practices for the management of geospatial resources are often relying on centralized architectures and outdated metadata formats for their enactment.

In order to allow researchers to maintain full control over the products they are providing, it is essential that the former start considering the management of their digital identities as a task that is complementary to the ordinary metadata management that is carried out in the provision of geospatial resources through geoportals. There are a number of tools and initiatives that can assist researchers in this activity and a growing corpus of freely available information that can support it, that is, the Linked Open Data cloud.

4.1. Acknowledgments

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

4.2. References

[1] European Commission 2007 Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) URL http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002

[2] European Commission 2008 Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata URL http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008R1205

[3] European Commission 2008 Corrigendum to Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (OJ L 326, 4.12.2008) URL http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008R1205R\%2802\%29

[4] International Organization for Standardization (TC 211) 2014 ISO 19115:2014 Geographic information – Metadata Tech. rep. URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=53798

[5] Dawes S S 2012 A Realistic Look at Open Data URL http://www.w3.org/2012/06/pmod/pmod2012_submission_38.pdf

[6] Bizer C 2011 Evolving the Web into a Global Data Space *Proceedings of the 28th British National Conference on Advances in Databases* BNCOD'11 (Berlin, Heidelberg: Springer-Verlag) pp 1–1 ISBN 978-3-642-24576-3 URL http://dl.acm.org/citation.cfm?id=2075914.2075915

[7] Haslhofer B and Klas W 2010 *ACM Computing Surveys* **42** 1–37 ISSN 03600300 URL http://dl.acm.org/citation.cfm?id=1667062.1667064

[8] Botts M and Robin A 2007 OpenGIS Sensor Model Language (SensorML) Implementation Specification Tech. rep.

[9] Botts M and Robin A 2014 OGC SensorML: Model and XML Encoding Standard Tech. rep.

[10] Cox S 2013 Geographic information Observations and measurements - OGC and ISO 19156 Tech. rep.

[11] Commissione di Coordinamento SPC 2013 Linee guida per linteroperabilità semantica attraverso i Linked Open Data URL http://archivio.digitpa.gov.it/notizie/linee-guida-open-data-interoperabili

[12] Oggioni A, Tagliolato P, Fugazza C, Pepe M and Carrara P 2015 Sensor metadata blueprints and computer-aided editing for disciplined SensorML *Proceedings of the 9th Symposium of the International Society for Digital Earth (ISDE)*

[13] Santoro M, Mazzetti P, Nativi S, Fugazza C, Granell C and Díaz L 2012 Methodologies for augmented discovery of geospatial resources *Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications* ed Díaz L, Granell C and Huerta J (IGI Global) chap 9, pp 172–203 ISBN 9781466609457 URL http://www.igi-global.com/chapter/methodologies-augmented-discovery-geospatial-resources/70448

[14] Fugazza C 2011 *Earth Science Informatics* **4** 225–239 ISSN 1865-0473 URL http://dx.doi.org/10.1007/s12145-011-0088-1