Beyond INSPIRE: an ontology for biodiversity metadata records

João Rocha da Silva[1], João Aguiar Castro[1], Cristina Ribeiro[2], João Honrado[3], Ângela Lomba[3], and João Gonçalves[3]

[1] Faculdade de Engenharia da Universidade do Porto/INESC TEC
joaorosilva, joaoaguiarcastro@gmail.com
[2] DEI—Faculdade de Engenharia da Universidade do Porto/INESC TEC
mcr@gmail.com
[3] CIBIO—Centro de Investigação em Biodiversidade e Recursos Genéticos
jhonrado, angelalomba@fc.up.pt, joaofgo@gmail.com

Abstract. Managing research data often requires the creation or reuse of specialised metadata schemas to satisfy the metadata requirements of each research group. Ontologies present several advantages over metadata schemas. In particular, they can be shared and improved upon more easily, providing the flexibility required to establish relationships between datasets and concepts from distinct domains. In this paper, we present a preliminary experiment on the use of ontologies for the description of biodiversity datasets. With a strong focus on the dynamics of individual species, species diversity, biological communities and ecosystems, the Predictive Ecology research group of CIBIO has adopted the INSPIRE European recommendation as the primary tool for metadata compliance across its research data description. We build upon this experience to model the BIOME ontology for the biodiversity domain. The ontology combines concepts from INSPIRE, matching them against the ones defined in the Dublin Core, FOAF and CERIF ontologies. Dendro, a prototype for collaborative data description, uses the ontology to provide an environment where biodiversity metadata records are available as Linked Open Data.

1 Introduction

Research data are often unique and valuable beyond the time frame of individual projects, but their long-term preservation depends on the existence of associated metadata records [5]. Many metadata schemas have been proposed in the past for different domains, often sharing elements with similar or matching meanings [8,11]. However, their combination or co-existence is not easy, often requiring translation or crosswalk operations [1].

Ontology languages are general-purpose knowledge representation technologies and can be adopted for capturing the nature of metadata records. They convey not only the syntactic rules that enforce the correctness of a metadata record but also the meaning of each descriptor used in a record—in a machine-processable way. They are essential for the description of resources on the Semantic Web [4]. The Linked Data principles

rely on a unique identifier, or URI, for each resource on the web, on the existence of links between those resources, and on the ability to retrieve resources linked to a particular one, all of them being represented using standard formats (e.g. RDF, OWL) and queryable via standard languages (e.g. SPARQL). Generalizing these guidelines to data that is publicly accessible yields the concept of Linked Open Data (LOD) [3].

Biodiversity information is rich and has many facets to take into account, from very specific ones such as taxonomies of species or community types, ecosystem process typologies or habitat descriptions, spatial and temporal resolution, to more general ones such as geo-referencing, institutional context, time of collection, methods and instruments applied to data collection, and the people involved in it [13].

We are focused here on the representation of biodiversity datasets, providing enough information to make them searchable and understandable by domain experts and relying on researchers to provide a minimal set of dataset-level descriptors. Moreover, we integrate biodiversity dataset description in Dendro [19,20], a prototype collaborative research data management platform. Dendro is ontology based, and takes advantage of the set of basic descriptors for biodiversity captured in BIOME, an ontology that includes the relevant properties of data collected by researchers focused on the patterns and dynamics of biodiversity and ecosystems. Conforming to previous work [18], we have incorporated in BIOME concepts from INSPIRE, an European Commission directive for representing geospatial metadata, to test its applicability in a data management and metadata creation scenario.

Using Dendro and the BIOME ontology, researchers can collaborate in the creation of metadata records that have all the advantages of a fully linked open data representation [20], including SPARQL querying and metadata record representation in formats suitable for LOD de-referencing (e.g. JSON, RDF or OWL).

2 About biodiversity data

Worldwide, environmental changes are acknowledged by their contribution to the increasing rates of biodiversity loss [7,17,18]. Understanding and assessing environmental and ecological change is thus essential if the baseline of natural capital is to be defined, in the context of monitoring schemes and long-term ecological research [17,18]. However, even though the production of spatially-explicit biodiversity data has been increasing, such assessment has been recurrently constrained by incomplete spatial and taxonomical indicator coverage [17].

Biodiversity data often refer to several taxonomic groups, community types, typologies of ecosystem processes, or habitat descriptions across distinct time periods, and are usually collected by distinct groups of researchers. Furthermore, assets of spatially-explicit data related to biodiversity have recently increased significantly, even if they present important differences regarding e.g. their source and methodological development, time of acquisition, and spatial resolution [15,18]. As a result, there is a pressing need towards the development of common and collaborative networks and tools towards the harmonisation and sharing of data from different sources and scales under common standards and languages [15,18].

Attempts to tackle such challenges have already been made in the context of specific research projects e.g. the European projects EBONE[1] and BIO_SOS[2], in the context

[1] http://www.wageningenur.nl/

[2] http://www.biosos.eu/

of broader initiatives e.g. the Group on Earth Observation Biodiversity Observation Network (GEO BON)[3]. Specifically, in the case of BIO_SOS project, a geo-portal has been developed so that all researchers from distinct partners and countries could share metadata concerning the datasets produced in the context of the project. The cornerstone of the implementation of this geo-portal was the definition of a metadata profile, built on the INSPIRE core profile, that each of the researchers would fill in during the data gathering stages. This then allowed the evaluation of the fitness of each dataset for the specific intended uses, using a methodology based on the metadata entries, that was itself embedded in the geo-portal [18]. This approach has been used by the CIBIO team in recent projects, as a standard toolkit for data management and quality assessment.

The INSPIRE directive 2007/2/EC[4] aims to create an European Union spatial data infrastructure, enabling the sharing of environmental information among public sector organisations to facilitate public access to spatial information across Europe [2].

The metadata elements specified in the INSPIRE recommendation include the identification of the resource, classification, geographic and temporal references, statements related to the quality and validity of the datasets, conformity with implementing rules on the interoperability of spatial data sets and services, constraints related to access and use, and the organisation responsible for the resource. Information about metadata records themselves is also necessary to ensure that records are kept up to date, and for identifying the organisation responsible for their creation and maintenance.

The INSPIRE directive also considers the possibility of users and systems providing a more detailed description of their resources. This allows them to use additional elements if these are prescribed by other international standards or working practices in their community.

3 The BIOME ontology

We now introduce BIOME, an ontology that reuses concepts from the INSPIRE directive 2007/2/EC[5], grouped according to the sections of the INSPIRE Geoportal Metadata Editor[6] and complemented by others specified in the BIO_SOS metadata quality guidelines [13].

INSPIRE specifies more than just the metadata representation and exchange formats, it also details how software infrastructures and web services must be designed, for example. Thus, building an ontology for the representation of the entire INSPIRE directive is a very complex task. This model in no way conflicts with other initiatives aimed at capturing the INSPIRE directive in a more fine-grained detail. This is one of the main advantages of modelling with ontologies, the ease with which different approaches can be combined and the incremental development of models at different granularity levels.

All the concepts created for this ontology were given annotations specifying their `rdf:label`s and `rdf:comment`s. This is important because a natural language description of the concepts is essential to ensure their interpretation by humans [6], and is a good ontology modelling practice overall. Moreover, it is actually necessary to specify

[3] `http://www.earthobservations.org`
[4] `http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008R1205`
[5] `http://inspire.ec.europa.eu/index.cfm/pageid/48`
[6] `http://inspire-geoportal.ec.europa.eu/editor/`

these annotation properties to use the ontologies with Dendro, as the platform uses their values to build its dynamic resource description interfaces.

The metadata elements recommended by INSPIRE comprise many categories, including information about the metadata records themselves. This way, metadata are kept up to date and the organisation responsible for their creation and maintenance can be easily identified. The directive also considers the possibility of users and systems providing a more detailed description of their resources. This allows them to use additional elements if these are required by other international standards or the working practices in their communities.

The BIOME ontology is designed for ease of use. It is simple enough to be both easily processable by machines and easily managed by data curators. We have minimized the number of object properties, focusing the model on some core classes and data properties and dispensing with constraints and axioms. The classes capture high-level concepts that represent both data and metadata, while the data properties correspond to the descriptors used in the description of the resources. Resources are therefore captured as instances of the defined classes.

Ontologies that follow this approach have been termed "lightweight ontologies". The Dublin Core (DC) ontology[7] and the Friend of a Friend (FOAF) are examples of two widely used lightweight ontologies [10]. We also promote the reuse of concepts from other ontologies, namely DC, FOAF and CERIF[8]. All the relations that we established between BIOME and the Dublin Core ontology were based on the "INSPIRE implementing rules for metadata and Dublin Core" [9].

3.1 Classes

The first step in the development of the BIOME ontology was to select the concepts to be represented as *classes*. The following concepts were considered essential to capture the top entities in the domain.

1. **Keyword**
   The *Keyword* class captures the concept of a keyword, a special term which can be associated to a resource; the keywords from a specific controlled vocabulary are examples of instances of the *Keyword* class.
2. **GeographicLocation**
   In BIOME this class is defined as a sub-class of the Dublin Core element *Location*, which is in turn a subclass of the *LocationPeriodAndJurisdiction* in Dublin Core as well.
3. **TemporalReference**
   This concept captures the temporal dimension of the data. The INSPIRE directive defines that at least a temporal reference must be provided.
4. **ResponsibleParty**
   The responsible party is an organisation responsible for the establishment, maintenance and distribution of the spatial data and services.

[7] a OWL version of the Dublin Core metadata schema can be found at `http://bloody-byte.net/rdf/dc_owl/`

[8] The Common European Research Information Format (CERIF) Ontology Specification provides basic concepts and properties for describing research information as semantic data. An OWL representation of CERIF is available at `http://www.eurocris.org/ontologies/cerif/1.3/`

Category	Property	Label
Identification	resourceTitle(*) resourceAbstract(*) linkage identifierNamespace identifierCode(*) resourceLanguage resourceLocator (≡ `dcterms:identifier`)	Resource title Resource abstract Linkage Identifier namespace Identifier code Resource language Resource locator
Classification	topicCategory* (⊑ `dcterms:subject`)	Topic Category
Keyword	keywordINSPIRE(*) keywordValue(*) (⊑ `dcterms:subject`) originatingControlledVocabulary: -title* -referenceDate* -dateType*	Keyword INSPIRE Keyword Value Originating controlled vocabulary: -title -reference date -data type
Geographic Location	geographicBoundingBox(*)	Geographic bounding box
Temporal reference	temporalExtentStartingDate(*) temporalExtentEndingDate(*) dateOfCreation* (≡ `dcterms:created`) dateOfLastRevision (≡ `dcterms:modified`) dateOfPublication (≡ `dcterms:published`)	Temporal extent starting date Temporal extent ending date Date of creation Date of last revision Date of publication
Quality & Validation	lineage	Lineage
Spatial Attributes	spatialRepresentationType spatialResolution spatialResolutionEquivalentScale spatialresolutionDistance	Spatial representation type Spatial resolution Spatial resolution equivalent scale (part of a spatial resolution record) Spatial resolution distance (part of a spatial resolution record)
Conformity	conformityDegree(*) specification conformityDate conformityDateType	Degree of conformity Specification Conformity date Conformity date type
Constraints	limitationsOnPublicAccess (⊑ `dcterms:accessRights`) conditionsApplyingToAccessAndUse (⊑ `dcterms:rights`)	Limitations on public access Conditions applied to access and use
Responsible party	organizationName(*) responsiblePartyEmail(*) responsibePartyRole(*)	Organisation Name Responsible party e-mail Responsible party role
Metadata on Metadata	organizationName(*) metadataPointOfContactEmail(*) metadataLanguage(*) metadataDate (⊑ `dcterms:date`)	Organisation name Metadata point of contact email Metadata language Metadata date

Table 1: Descriptors drawn or adapted from INSPIRE

Group	Property	Label
GeoCoMaS	projectName version diagnosticAndUsability	Project name Version Diagnostic and usability
ISO19115	distributionFormatName geographicExtentCode spatialRepresentation referenceSystem: -authority -identifier* -identifierCode* -identifierCodeSpace	Distribution format name Geographic extent code Spatial representation Reference system: -authority -identifier -identifier code -identifier code space

Table 2: Descriptors drawn from ISO 19115 and GeoCoMaS

3.2 Properties

Table 1 shows the concepts proposed for the BIOME ontology. It illustrates the categories that aggregate the descriptors under the heading **Group**, while presenting the **Properties** in each category—second column. The third column has the **Label** (`rdf:label`) for each property. The label facilitates the interpretation of the property name by users at the time of metadata creation in the data management platform interface. Table 2 shows the additional concepts that were drawn from GeoCoMaS[9] and ISO 19115 respectively. In the table, mandatory properties are marked with a $*$ symbol. A $-$ (dash) symbol before the property names in a sequence of properties indicates that the property immediately above them is used as a prefix; this is the case for *title*, *referenceDate* and *dateType* with respect to *originatingControlledVocabularyTitle*. In case there is a subsumption ($\sqsubseteq$) or equivalence ($\equiv$) relationship between a property in BIOME and another from a different ontology, the relationship is expressed in parentheses. For example, *resourceLocator* is equivalent to the `identifier` property in the DC ontology, so they are marked with ($\equiv$ `dcterms:identifier`) in their table line.

For the purpose of providing information about the metadata record itself, the properties *metadataLanguage*, *metadataDate* and *metadataPointOfContactEmail* were created. Resource identification is done using the properties *resourceTitle*, *resourceAbstract*, and *resourceLanguage* defined in the ontology. A *linkage* property was also created as a way to record the *resourceLocator* in the form of a URL, along with the *identifier*. The classification of spatial data and services can be described by a *topicCategory* property, that can assist users in finding a resource based on a topic search. The *Keyword* class can be described by a *keywordValue*, that assumes free-text values to represent "a commonly used word, formalised word or phrase to describe the subject"[10], contrasting with the *keywordINSPIRE*, that represents a INSPIRE thematic category, and whose valid values are found in the Gemet Thesaurus[11]. The *originatingControlledVocabulary* specifies valid *Keyword* values and must be "a formally registered thesaurus or a similar authoritative source of keywords"[12], as captured in the *title, referenceDate* and *dateType* of the corresponding controlled vocabulary. For representing the geographic location of the resource, we included the *geographicBoundBox* property, that must express the westbound and eastbound longitudes and the southbound and northbound latitudes of the intended location. On the other hand the *TemporalReference* class is described with values related to the *temporalEntextStarting/EndingDate*, the *dateOfCreation*, *dateOfPublication* and the *dateOfLastRevision*. We derive most of the temporal and geographical concepts from the DC classes *Location* and *PeriodOfTime*.

To address process history record-keeping and the overall quality of the dataset, along with the *spatialResolution* referring to the level of detail of a dataset, the *lineage*, *resolutionDistance*, and the *equivalentScale* properties were included. To register conformity of the dataset with implementing rules, or other specifications, we define the *conformityDegree* property to represent the degree of metadata conformity, the *specification* it conforms to, the *conformityDate* and the *conformityDateType*.

The constraints related to the access and use are stated via the *conditionsApplyingToAccessAndUse* and by the *limitationsOnPublicAccessAndUse*. Finally, the entity that

[9] GeoCoMaS is a normative system of good practices used by the CIBIO research team for managing files in their internal repository

[10] As described in the INSPIRE online documentation

[11] `http://www.eionet.europa.eu/gemet/en/themes`

[12] See footnote 10

holds responsibility over the resource must be described using the *organizationName*, the *responsiblePartyEmail* and a *responsiblePartyRole*.

To account for the possibility of a more detailed annotation, which is in compliance with the INSPIRE directive, the BIOME ontology also includes extra elements prescribed by the ISO 19115 [12]. These elements are intended to describe the spatial reference system used in the dataset (*referenceSystemCode*, *referenceSystemAuthority*, *referenceSystemIdentifier*, the digital mechanism used to represent spatial information (*spatialRepresentation*), the description of the geographic area through identifiers (*geographicExtentCode*) and the format in which the data is to be distributed (*distributionFormatName*).

The metadata records created by the researchers in this domain are complemented by elements from the GeoCoMaS fields, and must include a *version* of the resource, the identification of the project where the resource originated from (*projectName*), its domain expressed as a CERIF *Project*, which is in turn a subclass of the FOAF's *Project* class. A summary of data characteristics, quality and usability is specified in *diagnosticAndUsability*.

4 Ontology-driven data description

Dendro is a prototype data management platform currently under development at the University of Porto, built for LOD compliance from the ground up and complete with a SPARQL endpoint [19,20]. Its data model relies on the existence of publicly available ontologies (such as Dublin Core or FOAF) that can be loaded directly into the database layer of the system. As a result, the Dendro system is built on a fully graph-based model capable of representing every entity in the research data management environment, as well as the relationships between these entities, using explicit semantics [21]. The model can also be expanded by directly loading additional ontologies into the underlying graph database (using a SPARQL LOAD operation in Dendro's graph database interface), thus providing the flexibility that is expected of a cross-domain data platform. After loading an ontology, its properties become available as descriptors for files and folders.

Like a metadata schema, an ontology can be developed by a data curator, a data management role that may be assumed by either a domain expert or an information science expert. However, ontologies are advantageous to drive data description when compared with metadata schemas because they can be used to express not only the syntactic rules for metadata record compliance but also the semantics of metadata descriptors, in a machine-processable way. They can also be directly integrated into the data model, as an information representation and exchange format—in fact, Dendro cannot operate without first loading some ontologies that provide the metadata descriptors that allow the platform to work.

Figure 1 shows the different stages involved in the description of a dataset using Dendro. The BIOME ontology is based on the analysis of the INSPIRE directive, a relevant standard and the analysis of the metadata records that CIBIO researchers already produce (1). After loading BIOME into Dendro (2), its user interface (3) allows the creation of metadata records that include the same descriptors used in the original metadata sheets (see 1A and 3A), but with an ontology-based representation.

4.1 Dendro as a description platform for biodiversity datasets

Figure 2 shows a possible integration scenario between the current CIBIO information systems and Dendro. Data files and metadata continue to be created and loaded using

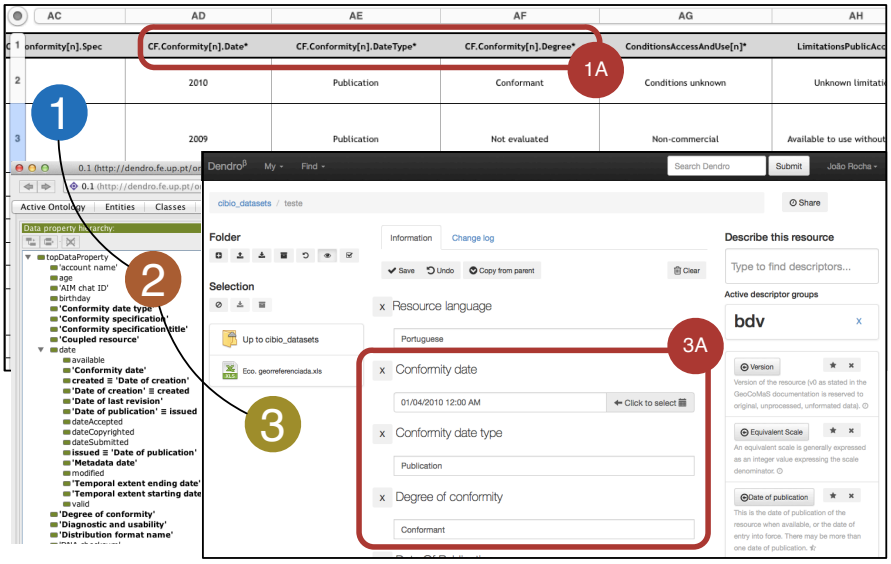

Fig. 1: Creating ontology-based metadata records in Dendro

the geo-portal, with Dendro being regarded as a project-centered data description platform. Metadata records in the geo-portal can be exported to Dendro for inclusion in a new project, initiating a process of collaborative description. At any time, the metadata records in Dendro can be imported back on the geo-portal, where they are queried and retrieved using INSPIRE-compatible web services.

5 Conclusions

In this paper we have presented BIOME, a lightweight ontology designed for the management of datasets in the biodiversity science domain. The research group is motivated to adopt metadata creation best practices, and already has a data repository in production, with domain-specific metadata conforming to the INSPIRE directive. Their goal its to couple the repository with a project-level collaborative platform for data organization and description. We have successfully loaded the BIOME lightweight ontology into Dendro, our prototype data management platform, and experimented with the descriptors which were already validated by the researchers in the biodiversity domain.

Working closely with the researchers, we build on the concepts defined by INSPIRE and combine them with others from general-purpose ontologies to create BIOME—an ontology that captures the metadata recommendations followed during the data description process. We promote interoperability by establishing relationships between the INSPIRE concepts and those defined in other widely used ontologies, while making annotations about all the concepts derived from the INSPIRE directive (using `rdf:label` and `rdf:comment`) and defining which properties are mandatory when creating a metadata record.

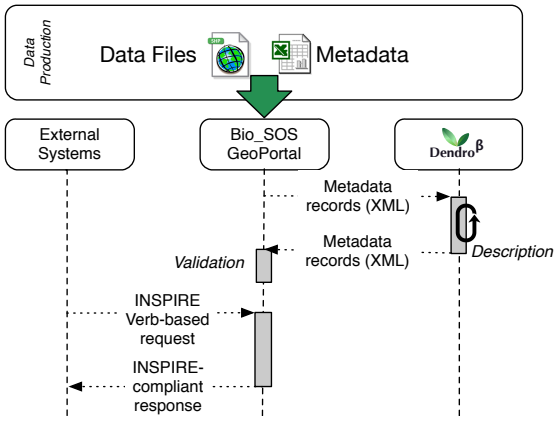

Fig. 2: Integrating the current data management platforms of the CIBIO team

The BIOME ontology focuses on establishing minimum requirements for managing biodiversity datasets, so the concepts and relationships were selected accordingly. More fine-grained ontologies are available, namely for describing specimen data (Darwin Core[13]), scientific observations (Extensible Observation Ontology or OBOE [16]) and other domain-specific concepts (Ecological Metadata Language or EML [14]). While still needing validation, this work is a first step in a path where, rather than starting with a very rich ontology, simpler ones are used in conjunction with a platform (Dendro) to manage datasets from their creation and minimal description is seen as a first stage in dataset preservation and sharing. Further feedback from the researchers as they start to manage their datasets will guide the next ontology and Dendro developments.

6 Acknowledgements

This work is supported by project NORTE-07-0124-FEDER000059, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). J. Rocha da Silva is also supported by research grant SFRH/BD/77092/2011, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). A. Lomba is supported by the Portuguese Science and Technology Foundation (FCT) through Post-Doctoral Grant SFRHBPD807472011. J. Honrado received support from FCT and COMPETE through project grant IND_CHANGE (PTDC/AAG-MAA/4539/2012).

References

1. M. Baca. Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage. *Cataloging & Classification Quarterly*, (April 2014):37–

[13] http://rs.tdwg.org/dwc/

41, 2003.
2. G. Bartha and S. Kocsis. Standardization of Geographic Data: The European INSPIRE directive. *European Journal of Geography*, pages 79–89, 2011.
3. T. Berners-Lee. Linked Data—Design Issues, 2006.
4. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.
5. C. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 2012.
6. D. Brickley and R. V. Guha. RDF Schema 1.1 W3C Recommendation, 2014.
7. S. Butchart, M. Walpole, et al. Global Biodiversity: Indicators of Recent Declines. *Science*, 328(5982):1164–1168, 2010.
8. LM Chan. Metadata Interoperability and Standardization – A Study of Methodology Part I. *D-Lib Magazine*, pages 1–34, 2006.
9. European Commission. State of progress in the development of guidelines to express elements of the INSPIRE metadata implementing rules using ISO15836. Technical report, European Commission, 2008.
10. O. Corcho. Ontology based document annotation: trends and open research problems. *International Journal of Metadata, Semantics and Ontologies*, 1(1):47–57, 2006.
11. M. Day. Integrating metadata schema registries with digital preservation systems to support interoperability: a proposal. *International Conference on Dublin Core and Metadata Applications (DC-2003)*, 2003.
12. International Organization for Standardization. ISO 19115:2003 Geography Information - Metadata. Technical report, 2003.
13. J. Honrado, J. Alonso, P. Castro, et al. The BIO_SOS metadata geoportal and the external quality of pre-existing datasets. Deliverable 4.5 of the FP7 project BIO_SOS. Appendix 1. Technical report, 2010.
14. M. B. Jones, C. Berkley, J. Bojilova, and M. Schildhauer. Managing scientific metadata. *IEEE INTERNET COMPUTING*, 5(5):59–68, October 2001.
15. A. Lomba, C. Guerra, J. Alonso, J. Honrado, R. Jongman, and D. McCracken. Mapping and monitoring High Nature Value farmlands: Challenges in European landscapes. *Journal of environmental management*, 143C:140–150, October 2014.
16. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296, October 2007.
17. J. Pereira, H. M., Brummitt, et al. Global biodiversity monitoring. *Frontiers in Ecology and the Environment*, 8:459–460, 2010.
18. I. Pôças, J. Gonçalves, B. Marcos, J. Alonso, P. Castro, and J. Honrado. Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring. *International Journal of Geographical Information Science*, (June):1–16, June 2014.
19. J. Rocha, J. Castro, C. Ribeiro, and J. C. Lopes. Dendro: collaborative research data management built on linked open data. *Proceedings of the 11th European Semantic Web Conference*, 2014.
20. J. Rocha, J. Castro, C. Ribeiro, and J. C. Lopes. The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In *Proceedings of the iPres 2014 Conference*, 2014.
21. J. Rocha, C. Ribeiro, and J. C. Lopes. Ontology-based multi-domain metadata for research data management using triple stores. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, 2014.