DOI 10.4430/bgta0051

Marine seismic metadata for an integrated European scale data infrastructure: the FP7 Geo-Seas project

P. Diviacco[1], R. Lowry[2] and D. Schaap[3]

[1] *Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Trieste, Italy*

[2] *British Oceanographic Data Centre, Liverpool, United Kingdom*

[3] *Mariene Informatie Service, Voorburg, The Nederlands*

(Received: May 19, 2011; accepted: September 15, 2011)

ABSTRACT The Geo-Seas EU 7th framework project aims to create a European scale open data space for marine geophysics and geology in which twenty-nine institutions will share their data assets. GeoSeas aims at interoperability with existing and well established data sharing initiatives in other fields of the marine sciences. At the same time it aims at being compliant with and stimulating the EU INSPIRE directive. Great attention was devoted to the development of the metadata model, which in this perspective is particularly critical. An implementation was chosen based on a three layered structure that leverages the possibilities offered by ISO19115, O&M and SensorML. The resulting model is demonstrated to be effective in solving both interoperability and domain specific issues while being light enough to avoid imposing an excessive burden on data managers.

Key words: metadata, ISO19115, observations and measurements, SensorML, Geo-Seas, geology, geophysics.

1. Introduction

Economic activities, sustainable environmental management and scientific research at regional, European and global scales all heavily rely on geological and geophysical data from the seabed and sub-seabed and on oceanographic data from the water column.

The Geo-Seas EU 7th framework project [Geo-Seas] is a 4 years effort gathering 26 marine institutions in 17 European maritime countries aiming at building a European data space in the field of marine geophysics and geology. This data space is to be integrated with the existing SeaDataNet [SeaDataNet] EU FP6th framework project initiative in the field of oceanography to create a synergy that will improve the location, access and delivery of a wide range of marine datasets and products to the designated research communities. This integrated data space takes interoperability with other international data initiatives, for example GeoSciML [GeoSciML], OneGeology [OneGeology] and Eurofleets [Eurofleets], into account and is commited to refer to and stimulate the EU Inspire [Inspire] directive through its developments.

In Europe, an important share of marine geological and geophysical observations is collected and analyzed by research institutions and national geological surveys. These often host substantial volumes of data collected by industry, government departments, academia and environmental organizations. Increasingly, these organisations are promoting the dissemination of their data by means of catalogues, supported in some cases by online data access facilities.

Despite well-established cooperation between national geological surveys and research institutes active in the European seas, it is currently not possible to federate separate national databases since these are built using different nomenclatures, reference levels, formats, scales and coordinate systems. This hampers direct integrated use of 'primary' data and the generation and dissemination of trans-boundary or multi-disciplinary datasets, products and services.

Examples of data types that can be delivered by Geo-Seas are seismics, bathymetry (including digital terrain models), lithology, mineralogy, geochemistry, sediment grain-size and geotechnical data.

The Geo-Seas vision comprises a European-wide data infrastructure, standardized practices by data centers, and middleware. Implementation of this is not trivial considering that twenty-nine institutions have to be coordinated, many data types have to be managed and integration with existing initiatives has to be taken into account.

The user experience in such a data space spans all steps from data discovery to data access and is supported by a mosaic of metadata models, data formats, vocabularies and services. In this paper we discuss the issues and solutions adopted in modelling metadata for marine exploration seismics.

2. An integrated data space

Marine studies are commonly multidisciplinary. Geo-Seas and SeaDataNet, albeit coming from different backgrounds, share the ambition to serve all the communities that focus on marine studies. This heads naturally to the concept of an integrated data space spanning all disciplines and data types.

SeaDataNet was originally designed to support data sharing in the oceanographic community, which of course uses data types that differ from those used by the geophysics and geology community. It is well known how different scientific communities living in different paradigms can diverge (Kuhn, 1962; Lakatos, 1970; Latour and Woolgar, 1979), so that the practices developed within one become incompatible with those developed in another.

The philosophy of science states that a given concept may exist only in the context or paradigm in which it is defined and only with difficulty can be that concept used elsewhere. As a result in any research field, but particularly in Earth sciences, there exists a spectrum of different visions of similar topics ranging from high abstraction levels (e.g., the research hypothesis) to lower levels (e.g., metadata parameters). The broader the community or the higher the level of abstraction the higher the level of divergence that can be seen. These observations can be attributed to the scientific process: namely the formulation of hypotheses.

This formulation of hypotheses is spread across all aspects of scientific research. The acquisition parameters of a geophysical survey, for instance, could be considered by a non-specialist as simply an administrative number. However to a specialist, they represent a piece of the bigger picture linking the sampling strategy to the hypothesis which the fieldwork is designed to test. Similarly, in the case of metadata, communities can be strongly divergent in the specification and creation of metadata to the detriment of the creation of strongly needed shared data spaces within that scientific community.

As SeaDataNet has already become a *de facto* reference in oceanography it was decided to

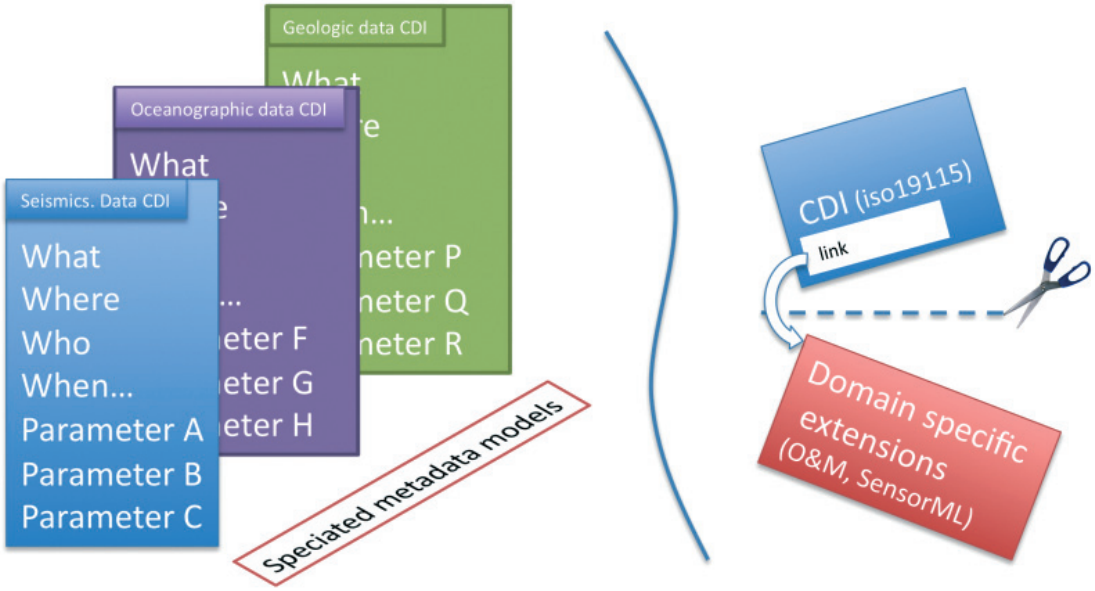


Fig. 1 - Two possible strategies: (left) add domain specific parameters to the original CDI structure, this would speciate the CDI, (right) separate domain specific parameters from a core discovery metadata structure represented by the CDI.

adopt its operational technologies as a starting point for Geo-Seas. However, in cases, such as seismic data, where they were not directly applicable, evolutions and extensions were developed. Metadata is a case in point, since the commitment was to balance a domain oriented solution while maintaining full compatibility with the SeaDataNet environment and at the same time referring to and stimulating the EU Inspire directive [Inspire].

Data discovery is the process by which an end user may know if the issues of interest can be addressed by existing data. Data should be catalogued and indexed based on metadata parameters that describe the characteristics of interest for end users to facilitate this.

SeaDataNet developed an ISO19115 [ISO19115] metadata profile and ISO19139 [ISO19139] XML encoding designed to describe an individual ‘deliverable data object’: a file or group of files containing the data from a single instance of a feature type such as a profile, point time series or trajectory. This is termed the Common Data Index (CDI) [CDI] and provides an insight into the availability and spatio-temporal coverage of marine data archived at the connected data centres.

Considerable effort went into the development of the CDI, including the development of tools [Mikado] and an infrastructure to manage standard mark-up terms in controlled common vocabularies (Latham *et al.*, 2009).

At the same time, potentially, each data type could be described through a specific set of parameters that would eventually result in discipline-specific search paths, practices, and metadata models.

This is not advisable for integrated cross-domain discovery of geophysical, geological and oceanographic data (Fig. 1).

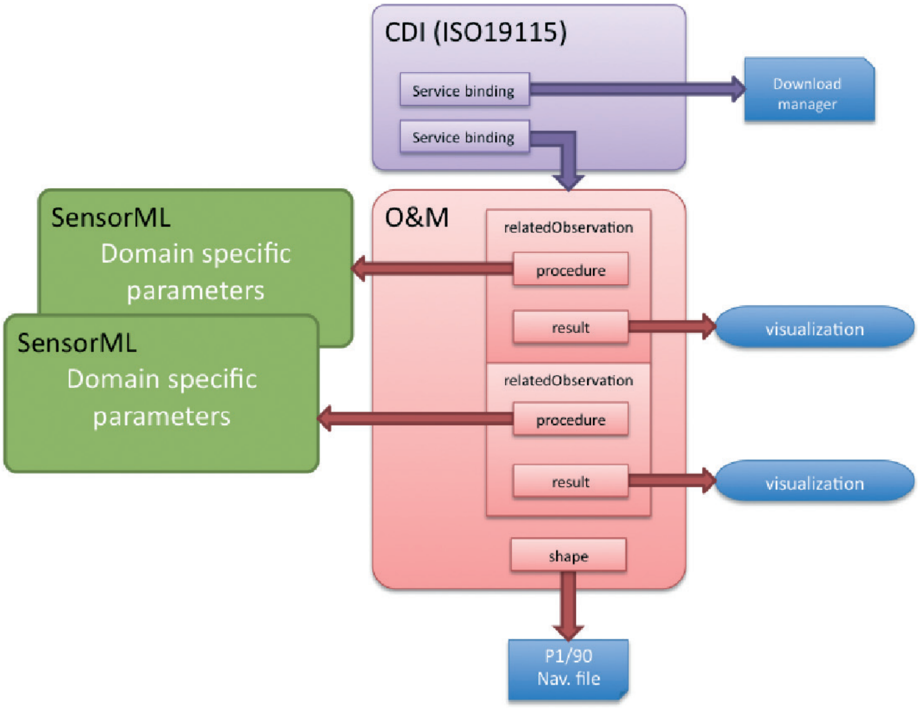


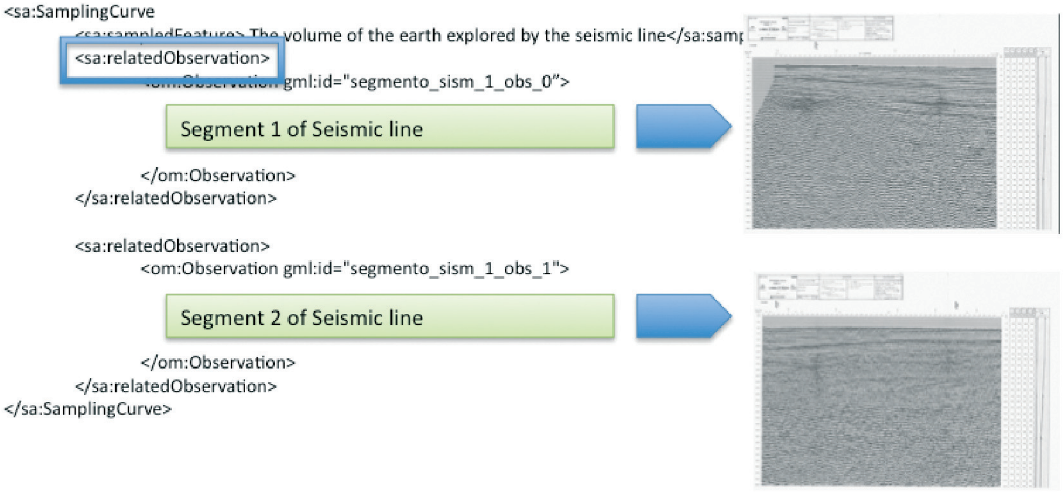
Fig. 2 - A scheme of Geo-Seas metadata model for seismic data. The ISO19115 compliant CDI representing the abstraction of a seismic line is linked via service binding to an O&M document. This allows for detachment of the seismic line from the actual files. The O&M document linking to one or multiple SensorML document allows the reporting of domain specific parameters, which can differ segment by segment.

To address this problem, it is useful to introduce a distinction between a core set of parameters shared by all data types, and domain-specific parameters. Metadata models based on ISO19115 [ISO19115] and Inspire [Inspire] favour core parameters as they are intended to foster cross-domain data sharing.

On the other hand domain-specific problems should be addressed with domain-specific tools. In the same way that there is no use looking for a screwdriver when a wrench is required, there is no use specifying the hydrophone streamer length amongst metadata parameters that describe current meter data. End users should be able to focus exclusively on their subset of interest: there is no advantage to an end user from data discovery if data of interest are irrelevant hits.

During the integration between Geo-Seas and SeaDataNet, this twofold nature of data discovery was built into the metadata structure for seismic data with an upper ISO19115 generic layer coexisting with another deeper layer, based on Observations and Measurements (O&M) [O&M] and SensorML [SensorML], addressing domain-specific issues. O&M is an Open Geospatial Consortium (OGC) [OGC] open standard conceptual model with an XML encoding based on Geographical Markup Language (GML: ISO19136 [ISO19136]) that is designed to describe data collection activities with the potential to solve the above mentioned issues. SensorML is an XML based OGC standard to exchange information focusing on sensors and sensor systems.

The CDI was adopted as the upper discovery layer, therefore facilitating integration between



```
<sa:SamplingCurve
        <sa:sampledFeature> The volume of the earth explored by the seismic line</sa:samp
        <sa:relatedObservation>
                <om:Observation gml:id="segmento_sism_1_obs_0">
                </om:Observation>
        </sa:relatedObservation>

        <sa:relatedObservation>
                <om:Observation gml:id="segmento_sism_1_obs_1">
                </om:Observation>
        </sa:relatedObservation>
</sa:SamplingCurve>
```

Fig. 3 - The <sa:relatedObservation> element allows several segments of the same seismic section to be gathered together.

Geo-Seas and SeasDataNet while at the same time highlighting its ISO19115 vocation (Fig. 2).

3. Domain specific issues

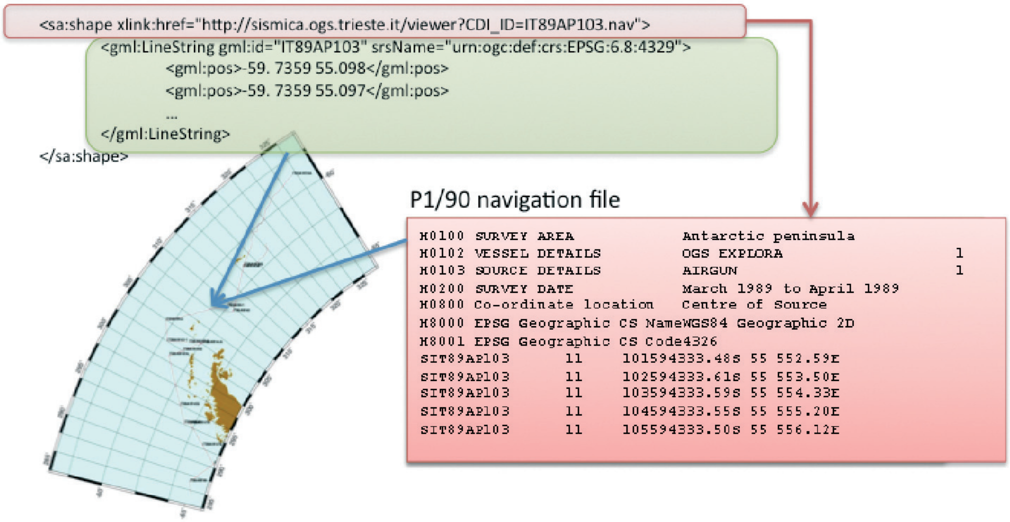
Domain specific problems for seismic data include the specification of the feature for the ‘deliverable data object’, data licensing, interoperability with other European and international initiatives, and the definition of a metadata parameter list that makes sense yet is light enough for project partners, especially from academia, who do not specialize in data management.

One of the first issues considered was the definition of the entity to be described by a CDI record. While the community naturally identifies this entity as the concept of the seismic line, technology legacies indicate the need to handle seismic line segmentation. For example, in Seiscan/SeiscanX EU projects (Miles *et al*., 1997) five geological research institutes (all partners in Geo-Seas) rescued and scanned large amounts of old seismic reflection data from deteriorating paper records, which were often segments of a whole seismic line.

A solution was developed that addressed both the community and tehnological issues. This was to build a domain-specific metadata layer beneath the CDI record (Fig. 2) as external XML documents linked to the CDI record through a service binding. This is simply a URL accompanied by an encoding to inform a client of the nature of the document or service to which it points. This layer provides both a home for domain-specific information and a mechanism for implementing a one-to-many relationship between a CDI record and multiple deliverable data object components.

It is built from an O&M XML document which itself links to one or more SensorML XML documents.

For example, the issue of addressing multiple segments within a seismic line is solved in O&M by the abstraction of the single seismic line, represented by a CDI instance, into multiple



```
<sa:shape xlink:href="http://sismica.ogs.trieste.it/viewer?CDI_ID=IT89AP103.nav">
    <gml:LineString gml:id="IT89AP103" srsName="urn:ogc:def:crs:EPSG:6.8:4329">
            <gml:pos>-59. 7359 55.098</gml:pos>
            <gml:pos>-59. 7359 55.097</gml:pos>
            ...
    </gml:LineString>
</sa:shape>
```

P1/90 navigation file

```
H0100 SURVEY AREA             Antarctic peninsula
H0102 VESSEL DETAILS          OGS EXPLORA                    1
H0103 SOURCE DETAILS          AIRGUN                         1
H0200 SURVEY DATE             March 1989 to April 1989
H0800 Co-ordinate location    Centre of Source
H8000 EPSG Geographic CS NameWGS84 Geographic 2D
H8001 EPSG Geographic CS Code4326
SIT89AP103      11     101594333.48S 55 552.59E
SIT89AP103      11     102594333.61S 55 553.50E
SIT89AP103      11     103594333.59S 55 554.33E
SIT89AP103      11     104594333.55S 55 555.20E
SIT89AP103      11     105594333.50S 55 556.12E
```

Fig. 4 - Management of positioning. Through the <sa:shape> element it is possible to store positioning and link to an external P1/90 navigation file with shot/cdp indexing.

objects through <sa:relatedObservation> elements, each of which refers to a different segment of the same seismic line (Fig. 3).

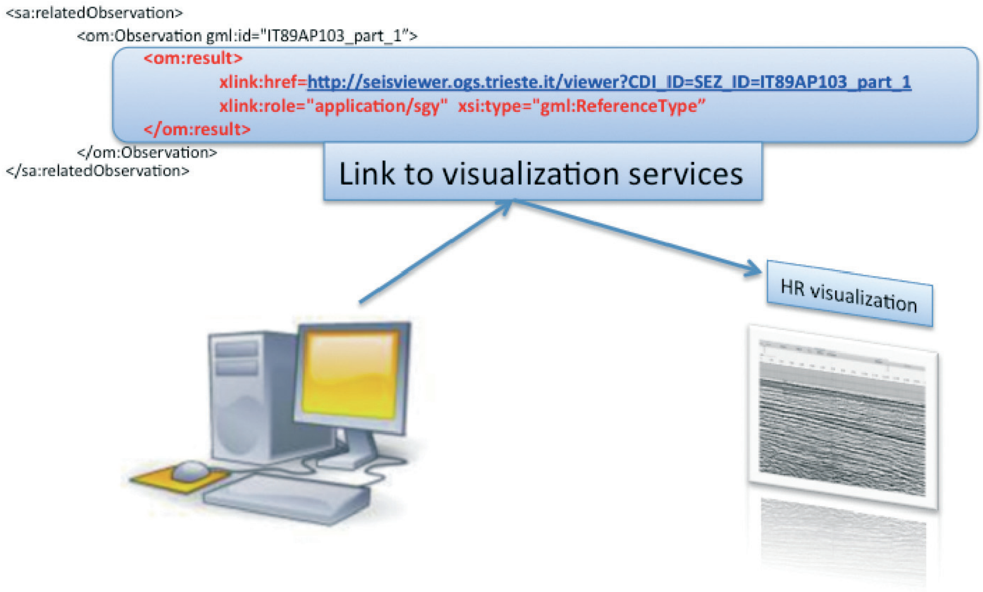
Another very important problem is the handling of seismic-specific spatial information, since the association between shot/cdp, traces and geographic location cannot be tackled using GML grammar. Marine seismic data positioning is traditionally handled via a navigation file as U.K.O.O.A. P1/90 [P1/90], that allows geo-referencing at any granularity and addressing of seismics-specific events. As this tool is well-proven, Geo-Seas will continue its use by embedding references in its metadata model, utilizing the <sa:shape> element provided by O&M (Fig. 4).

This can hold both a link to a URL, where an U.K.O.O.A. P1/90 navigation file can be exposed, and a GML linestring element representing the track.

4. Data preview, licensing and access

Access conditions for seismic data is an important concern in the case of scientific research. In the oil exploration and production industry there is long experience in data brokering and many initiatives have established stable and efficient means to find, buy and sell data. However in the field of scientific research this experience cannot be directly applied as the sociological and economic dynamics of the research community are rather different.

In both the commercial Exploration and Production (E&P), and scientific communities it is possible to distinguish two classes of actors: data owners and data seekers/users. However, while in the E&P industry the focus is on the unidirectional process of selling data, in the case of scientific research the process is bidirectional. Data are also used to attract external researchers and projects with the goal of positioning the data owner within the scientific community. These dynamics (Latour and Woolgar, 1979; Whitley, 2000; Diviacco, 2007, 2008, 2012) are quite complex because the actual process of science production and the sociology of science are



```
<sa:relatedObservation>
        <om:Observation gml:id="IT89AP103_part_1”>
            <om:result>
                    xlink:href=http://seisviewer.ogs.trieste.it/viewer?CDI_ID=SEZ_ID=IT89AP103_part_1
                    xlink:role="application/sgy"  xsi:type="gml:ReferenceType”
            </om:result>
        </om:Observation>
</sa:relatedObservation>
```

Fig. 5 - Service binding to external visualization facilities.

mutually conditioned.

The fundamental aspect, of these dynamics is that data owners are faced with two problems: the need to open their data to attract new collaborations, while at the same time the need to protect them. If data were fully public and downloadable, users could be tempted not to involve the data owner in their exploitation. A new paradigm is required to relieve the anxiety of data owners, while supporting the needs of a scientific community with ever growing data needs.

To address these issues, some work has already been developed towards a web based, server side data access that would minimize the need of data downloading (Diviacco, 2005, 2012). The fundamental criterion is that data can be visualized at resolutions depending on the licensing policy, allowing preview of low resolution and watermarked data to all registered users. Upon data selection, a form of agreement between the user and the data owner would set the entitlement to access the data, which can be granted by an interactive server side visualization facility limited to stacked data with some basic seismic data processing tools. In essence, the key feature of this paradigm is that data are not copied outside the scope of the data owner IT framework, which prevents loss of control of data (Fig. 5).

To embed this paradigm within Geo-Seas there should be ways to link data access facilities from within metadata. Since this is a domain-specific request, it will take place outside the CDI, and specifically within the O&M XML document. The O&M specifications offer an element named <om:result> that can host a web link to a visualization service, which would then handle access to that specific data. If there are multiple segments of a seismic section each of them could be accessed separately as this element is contained within the <sa:relatedObervation> element.

Data owners must provide a visualization facility that can handle data access as mentioned above. Geo-Seas is currently developing a tool that will be deployed to all partners in order to

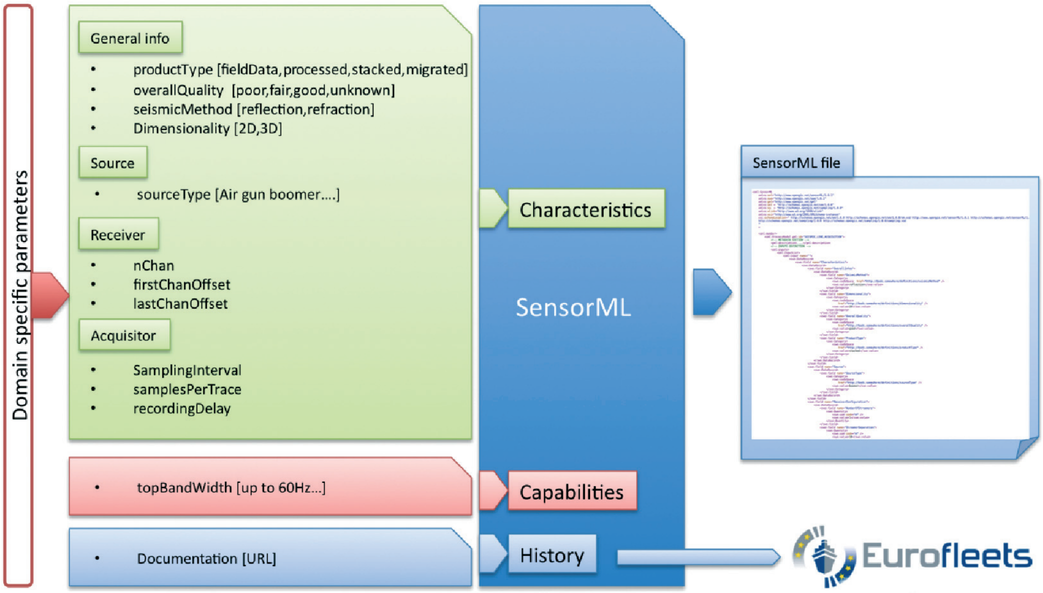


Fig. 6 - The light but focused model for domain specific parameters held in the SensorML section. Parameters are classified in three sections: (I) Characteristics (II) Capabilities (III) History.

offer this functionality.

Alternatively, if the licensing policy of the data owner permits or if there are specific agreements between data owner and user, data could be downloaded using the facilities already operative within SeaDataNet. These are based on the metaphor of a shopping basket, similar to those that can be found in most of the e-Commerce systems.

5. Domain specific parameters

Parameters such as sampling interval, streamer length or trace length provide fundamental information on the possible usage of that particular dataset. Identification of unsuitable datasets at the discovery stage avoids the waste of an enormous amount of time loading and checking data. A good metadata model needs to balance two contrasting forces: on one side we have the end user, who requires as many parameters as possible to perform useful queries; on the other side we have data owners, who need to populate records with these parameters. Within Geo-Seas many partners are scientific institutions without a data management unit. For these organizations highly complex metadata models can become an obstacle that prevents information being entered in the system. In some cases, some information is not available. A light, but meaningful, domain-specific parameter model is therefore needed.

Within the GeoSeas model, this function is accomplished by the SensorML document. SensorML is built of standard models and XML schemas for describing sensor systems and processes associated with sensor observations, so that it can be considered complementary to O&M. SensorML offers a Metadata Group with several sections. In Geo-Seas we use three of these to provide a light metadata model (Fig. 6):

I) Characteristics, describing specific parameters of the acquisition device;

II) Capabilities, describing the possible applications of data;

III) History, describing what happened during data acquisition, and possibly also afterwards.

The Characteristics section holds information such as source or streamer characteristics, sampling interval or trace length. The Capabilities section was linked to a vocabulary that states the top resolution (derived automatically from sampling interval) of the data. The History section has been implemented as a link to an external document where information on what happened during data acquisition may be written

However, this usage of the History section does not exploit its full potential. A much better solution would be to link to structured documentation, such as XML eventlist, rather than plaintext. Such a document is being developed as part of an automatic event logging and reporting tool within another FP7 EU project named Eurofleets.

6. Conclusions

Geo-Seas has developed a metadata model which aims to address domain specific issues and enrich ISO19115 metadata with scientifically-relevant content while at the same time providing interoperability with initiatives in other scientific disciplines.

This has been achieved through a three-layered model in which O&M acts as a bridge between cross-domain discovery relying on ISO19115 and domain specific parameters coded in SensorML.

This approach was implemented for the case of seismic data but in the long term it is proposed to provide a generalized solution applicable to any data type. The alternative to this is the proliferation of hundreds of domain-specific models that then need to be made to interoperate.

Aknowledgements. We would like to thank Alessandro Busato, Marc Shaming, Peter Miles, Oliver Clements, Arjen de Korte and Nigel Wardell for their contribution to this work.

REFERENCES

Diviacco P.; 2005: *An open source, web based, simple solution for seismic data dissemination and collaborative research.* Computer & Geoscience, **31**, 599-605.

Diviacco P.; 2007: *Data systems and the social aspects of scientific research.* In: Proceeding of the European Geosciences Union (EGU), Vienna, IGO 2007-A-02542, GI10-1FR4P-0328, poster.

Diviacco P.; 2008: *Col-laboratories, tools to foster the collaborative attitude within a scientific community while protecting data assets.* In: Proc. 33° International Geological Congress, Oslo, August 6-14, 2008.

Diviacco P.; 2012*: Towards a collaborative research data space in Geophysics.* Mediterranean Marine Sciences, in press.

Kuhn T.S.; 1962: *The structure of scientific revolutions*. University of Chicago Press, Chicago, U.S.A., 173 pp.

Lakatos I.; 1970: *Falsification and the methodology of scientific research programmes.* In: Lakatos I. and Musgrave A. (eds), Criticism and the growth of knowledge, Cambridge University Press, U.K., pp. 91-196.

Latham S.E., Cramer R., Grant M., Kershaw P., Lawrence B.N., Lowry R., Lowe D., O’Neill K., Miller P., Pascoe S., Pritchard M., Snaith H. and Woolf A.; 2009: *The NERC DataGrid services.* Philosophical Transactions of the Royal

Society A-Mathematical Physical and Engineering Sciences, **367**(1890), 1015-1019. doi:10.1098/rsta.2008.0238.

Latour B. and Woolgar S.; 1979: *Laboratory life: the construction of scientific facts.* Princeton Un. Press, Princeton NJ, USA, 294 pp.

Miles P.R., Schaming M., Casas A., Sachpazi M. and Marchetti A.; 1997: *Capturing a European legacy.* EOS Trans. AGU, **78**, 582.

Whitley R.; 2000: *The intellectual and social organization of the sciences.* Clarendon Press, Oxford, 319 pp.

WEB REFERENCES

[CDI]	http://www.seadatanet.org/Data-Access/Common-Data-Index-CDI
[Eurofleets]	http://www.eurofleets.eu/
[Geo-Seas]	http://www.geo-seas.eu/
[GeoSciML]	http://www.geosciml.org/
[ISO19115]	http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
[ISO19139]	http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557
[ISO19136]	http://www.iso.org/iso/catalogue_detail.htm?csnumber=32554
[Inspire]	http://inspire.jrc.ec.europa.eu/
[O&M]	http://www.opengeospatial.org/standards/om
[OGC]	http://www.opengeospatial.org/
[OneGeology]	http://www.onegeology.org/
[P1/90]	http://www.epsg.org/exchange/p1.pdf
[SeaDataNet]	http://www.seadatanet.org/
[MIKADO]	http://www.seadatanet.org/Standards-Software/Software/MIKADO

Corresponding author: Paolo Diviacco
Istituto Nazionale di Oceoangrafia e di Geofisica Sperimentale (OGS)
Borgo Grotta Gigante 42/c, 34010 Sgonico, Trieste, Italy
Phone: +39 040 2140380; fax: +39 040 2140438; e-mail: pdiviacco@ogs.trieste.it