Bollettino di Geofisica Teorica ed Applicata Vol. 57, n. 1, pp. 59-69; March 2016
DOI 10.4430/bgta0163

A metadata model for cross-domain marine data management, the SeaDataNet/Geo-Seas experience

P. Diviacco[1], R. Casas Munoz[2], J. Sorribas[2] and T. Loubrieu[3]

[1] *Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS), Trieste, Italy*
[2] *Consejo Superior de Investigaciones Científicas, Barcelona, Spain*
[3] *IFREMER, Plouzane, France*

(Received: May 28, 2015; accepted: October 1, 2015)

ABSTRACT Data management in the marine sciences faces the complex issue of addressing contrasting cognitive models across scientific domains, meaning the various representations that different communities build for overlapping areas of interest. Each of these communities constructs an identity on a specific and personal set of inherited backgrounds, practices, and tacit knowledge, which are mirrored in how they understand the targets of their studies. From a practical point of view, this can result in mis-linking observations and usage. This paper reports on the work done within the EU FP7 SeaDataNet project to tackle such problems through an integrated discovery and data access paradigm. It is based on a flexible metadata model that allows researchers to link domain-specific metadata profiles encoded using SensorML OGC standard, the general discovery framework based on an ISO 19115/19139 profile called CDI, and the data through a common hub based on the Observations and Measurements OGC standard.

Key words: marine data management, metadata model, observation and Measurements, Sensor Web Enablement, data discovery and access.

1. Introduction

In environments such as the sea, physical, chemical, and biological phenomena are strongly interlaced. Only very seldom can these complex systems be studied by isolating one parameter from the other. This often means that scientists from different backgrounds and disciplines work together, analyzing from one point of view, observations that could have been acquired under different perspectives. Traditional ways of using data are questioned, while new approaches might arise. Data are joined from different domains, creating new information spaces, meaning, initiatives, activities, and IT systems able to collect and manage data.

In addition to this, observation and measurements at sea involve the deployment on the same vessel of multiple instruments from various domains. Considering the costs of such sophisticated equipments, it is very important to carefully manage the access to instrumentation in order to optimize its use. Teams of scientists from different backgrounds need to collaborate, from observation planning to acquisition at sea to data sharing, potentially raising conflicts in cognitive models of knowledge that absolutely need to be addressed.

In exploring such new domains and spaces, scientists need some kind of "compass" to support their work, helping them to avoid miscommunication, and misuse of resources.

Within this perspective, two very important activities can be identified: data discovery and data access.

Data discovery is the activity (including the tools and the facilities) that allows a scientist to find the specific data of interest. Data access is the activity (including the tools and facilities) that, once data are identified, allows scientists to use it, be it local work after downloading of the data, or remote work, generally using web-based systems.

Data discovery has so far been the domain of metadata. Metadata is data about the data. It can be compiled either manually or automatically. The former takes a lot of time, and few institutions can afford the specialized personnel to take care of filling in such information. Automatic metadata extraction, instead, is generally based on calculations performed on the data themselves, such as number of observations, samples, or size of data.

Metadata can be generic information, such as location or timing, or it can be domain-dependent parameters such as seismic data sampling rate or sea surface temperature recording time range. A distinction is often made when searching for data using these two types of metadata between (proper) "discovery", where generic metadata are used, and "browsing", where domain-specific metadata are used. These latter ones are particularly interesting because they are linked to the observation strategy and therefore prone to be understood differently in different communities.

In the quest for data, therefore, it is difficult to manage such semantic-rich domain-specific content, since it is necessary to simultaneously consider the multiple paths employed in the search; one for each scientific community's point of view or paradigm.

We will, within the scope of this paper, categorize (proper) discovery and browsing under the generic term of data discovery.

2. The gap between discovery and access

Traditionally, data discovery and data access have been handled separately. Even the skills of people involved in these activities have often been different: data discovery has been the domain of semantics and databases, while data access has been focused more on space/time-varying analysis.

In contrast with this, a new trend (Diviacco and Busato, 2013; Baumann, 2014) suggests that the barriers between these two allegedly different worlds should be removed and that they can be handled within a single vision.

A practical example of how this can be put into practice is the selection of data on the basis of its ability to address issues such as the presence of noise or the availability of information on sensor calibration.

In the first case, data can be contaminated by the effects of phenomena that within one context are considered noise but within another provide a useful signal.

An example of this is the recent interest in surface waves in seismic prospecting. For deeper prospecting, they generally are cut off, but, on the contrary, inversion of such arrivals provides important information on the most shallow layers of the Earth surface (Forbriger, 2003). It is

important therefore that, upon data selection, the actual data can be previewed in order to reveal such signals.

Regarding calibration in the field of oceanography, various instruments have been used for decades to measure ocean physics. They have different features and in particular different levels of precision. For example, the computation of depth associated with each temperature for an expendable Bathy-Thermographs (XBT), widely used for the measurement of the ocean's upper layers, depends on an equation which converts the elapsed time since the probe entered the water to a depth value. In the past, different equations have been applied to different types of probes with variable success (Hamon *et al.*, 2011). In domains such as climatological studies and in particular water heat content measurement, where this information is crucial, the quality of each temperature profile is linked to the provenance information describing the instrument used and the equations applied for the computation of the coordinates and phenomena values.

Integration of data selection based upon metadata and data access is needed for other reasons. If a metadata-only approach is used, each discipline will look for its specific features of interest (FOIs), which may not be particularly relevant for others. At the same time, it is also important to highlight that data discovery based on data access only could be a burdensome process, since new users will start data analysis each time from scratch. Metadata, on the contrary, have the advantage of being usable throughout successive runs.

Traditionally, data selection has mostly been based on metadata, but we have demonstrated that both metadata and data access have advantages and disadvantages. We contend that a new perspective should be followed where these two approaches are integrated and available to users.

3. An integrated discovery and access metadata model

To address these issues, a good compromise could be a mixed approach where data discovery/browsing based on metadata and preview of the data is flexible, depending on the specific needs of a project. End users could filter among large hits of possibly interesting data sets, based upon metadata parameters that identify those data sets that theoretically better match the request. The selected subset can later be previewed using a URL, embedded in the <om:result> element. Then, it would be possible to see if the selected subset matches the expectations, such as, for example, quality or the presence of possible domain-specific features. In the case that the subsets do not, they can be filtered out, so that, eventually, the end user will be directed only to the data of potential interest. This usually saves a lot of time that would be otherwise wasted in downloading, uploading, and processing irrelevant data.

4. Sensor Web Enablement (SWE)

To allow the integration between discovery and data access, it is necessary to devise a metadata model that connects metadata with observations, while at the same time preserving the ability to cross disciplinary domains and paradigms.

To allow this, several tools and Open Geospatial Consortium (OGC) approved standards are being adopted within the marine research community. We will focus on the application of two of them:

- the Sensor Model Language (SensorML), which provides standard models and XML encoding for describing processes and processing components associated with the measurement and post-measurement transformation of observations. SensorML provides a robust means for defining the physical characteristics and functional capabilities of physical processes of sensors and actuators;
- the Observations & Measurement (O&M) standard, which defines a conceptual model and an XML implementation of schemas for observation results and for features involved in sampling when making observations.

O&M and SensorML are among the pillars of the Sensor Web Enablement (SWE) standards — a suite of standards that enable all types of sensors and data infrastructures to become discoverable and accessible via the web.

We will report on how we developed O&M and SensorML metadata schemas within several EU projects, where users have been allowed to browse and access the same data sets from different disciplines or domains, extending the range of possible users and improving the visibility of data.

5. Data management infrastructures

5.1. The SeaDataNet infrastructure

The SeaDataNet Project is a joint effort of 44 scientific institutions active in monitoring the marine environment in the European area and preserving marine observations long term by creating a network among the IODES's European National Ocean Data Centres (NODC).

For long-term data preservation, there is a requirement for high-quality data and metadata which demands strenuous efforts at NODCs to collect, harmonize, and standardize contributions from operators of observatories. This causes a mean time lag from observation time to data publication in the SeaDataNet infrastructure of several years. In the meantime, several observation infrastructures have organized themselves to coordinate the platform maintenance and deployment and harmonize the data management with their own dedicated standards. For operational oceanography, for example, the ARGO program (http://www.argo.net/) provides near real-time publication of their metadata and data through their global data assembly centres (GDAC).

To ease the data and metadata collection process at NODCs, to bridge the information flow from GDACs to SeaDataNet infrastructure, and to reduce the delay in the publication of observations, while not compromising the multidisciplinary coverage of SeaDataNet, the Sensor Web Enablement standards have been chosen. Specifically, SensorML has been employed for observatories' descriptions, while O&M has been used for observation data and metadata. Eventually, a common and flexible language will be used and understood by observatory operator systems and marine data management systems.

SWE standards have demonstrated their flexibility and suitability for sensor information management beyond the marine community. Within the marine research community, profile specifications and tools are required to produce and ingest marine data and metadata in SeaDataNet systems. This initiative plans to target the observatory operators (sometimes named PI for primary investigators) as users who will produce this information and also get services such as real-time alerts or publication tools in return.

SeaDataNet is thus defining a common unified profile for marine observations and specific templates dedicated to instruments (for example mooring, rosette with Niskin bottle and CTD probe, thermosalinograph or gravimeters on board research vessel, and others). The profile will be used within the range of tools that SeaDataNet is developing so that observatory operators can manage their metadata and data in their own online workspace, sometimes called in the cloud.

5.2. The Geo-Seas Project

The Geo-Seas FP7 EU project is a sibling of the SeaDataNet project focused on Geology and Geophysics. It also takes into account the experience and developments emerging from international projects, such as OneGeology and GeoSciML, that are oriented towards the management of geological data and where many of the Geo-Seas partners are also partners.

Besides the availability of a large number of data sets and observations in these fields, Geo-Seas shares with SeaDataNet several basic technologies and capabilities, while aiming to extend what was previously available in the field of data management in terms of crossing the divide between data discovery and data access.

5.3. A possible cross-domain metadata model

Within the Geo-Seas EU project, a metadata model has been developed (Diviacco *et al.*, 2012) that SeaDataNet wishes to adopt, adapt, and extend in order to obtain a cross-domain metadata model.

The Geo-Seas metadata model is strongly centered on a core Observation and Measurement – O&M (OGC and ISO19101) metadata layer (in blue in Fig. 1), which allows an identification and discovery layer (in the SeaDataNet and Geo-Seas case an ISO 19139/19115 profile named Common Data Index-CDI) to be detached but at the same time linked to a multiple, domain-specific, metadata profile. This approach is summarized in Fig. 1 where on the left a blue node identifies the CDI XML file that reports generic information such as what, where, or who is associated with a data set without providing detailed domain-specific information [for further details on the CDI format, please refer to Schaap and Lowry (2010)]. The green cluster of elements is related to the sampling strategy implemented during the data acquisition. This is necessary every time the observation is mediated, which is almost always. In the oceanographic, environmental, or geologic domain, sampling strategies deal with geospatial location and therefore need to explicitly name a feature type. This can be one of those listed in the O&M part2 sampling documentation, such as profile, swath, or station, while positioning itself can be stored in the shape element using the gml format, e.g.,

```
<gml:Curve gml:id="ttt-1" srsDimension="2" srsName="EPSG:4326">
    <gml:segments>
        <gml:LineStringSegment>
            <gml:posList>13.30425 43.956861 .... 13.317444 43.965 </gml:posList>
        </gml:LineStringSegment>
    </gml:segments>
</gml:Curve>
```

or using a link to a positioning file, e.g.,

```
<sa:shape xlink:href="http://diam04.ogs.trieste.it/Geo-Seas/B-401/B-401.uko"/>
```

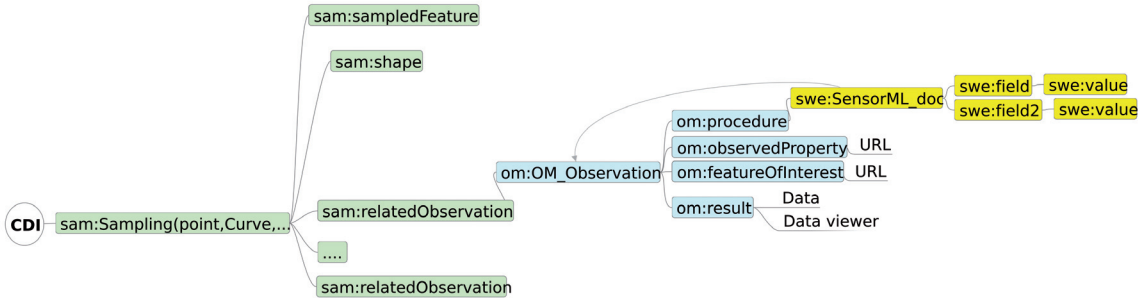

Fig. 1 - GeoSeas metadata model updated and tuned for the needs of SeaDataNet.

The sampling approach allows, using the sam:relatedObservation element, aggregation of multiple data sets under the same observation response. This can be very useful where multiple versions of the same data set are available or where a data set has been acquired as multiple segments. Each of these related observations should then wrap the O&M that contains the SensorML XML reference pertaining to that observation.

O&M is hard-typed, meaning that there are already prepared elements that account for some functionalities. SensorML, on the contrary, is soft-typed, allowing greater flexibility in the use of its elements. A very important synergy emerges then between O&M and SensorML, where the latter can be shaped to fill the rigid O&M elements.

The O&M <om:procedure> element, for example, is devoted to the description of observation strategies. This element can be used to report domain-oriented browsing metadata that specify the conditions under which data have been acquired. These can later be used to select data according to specific end-user needs. An example of this would be providing the length of a recording and thus enabling the end user the ability to understand if what he/she is looking for falls within the specific range he/she needs.

The joint use of O&M and SensorML allows association of an <om:procedure> with different, possibly domain-specific, SensorML extensions. These can range from a description of the sampling strategy held in a controlled vocabulary term up to a content-rich SensorML document where detailed information on acquisition and processing can be encoded.

A very useful strategy to use in providing values for the <om:procedure> element is to use URLs. Considering the cases of the link to a vocabulary term, this can be coded as in the example below:

```
<om:procedure xlink:href=”xlink:href=http://www.utm.csic.es/sos/kvp?
request=DescribeSensor&procedure=ID_29SG_TERMOSALINOMETERFLUOROMETER_
SYSTEM”/>
```

Considering the case of a web link to a SensorML document, this can be coded as follows:

```
<om:procedure xlink:href=”http://diam04.ogs.trieste.it/Geo-Seas/B-401/Linea_B-401_sml.xml”/>
```

A SensorML document linked by the <om:procedure> element can be very detailed and can contain detailed domain-specific information. Diviacco *et al.* (2012) reports on the case of seismic

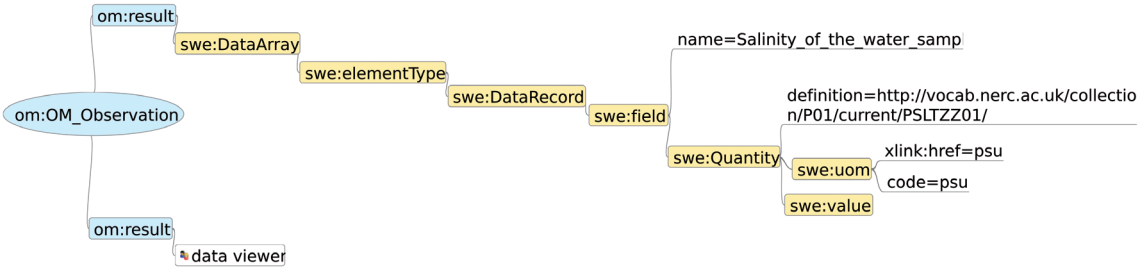

Fig. 2 - Example of the joint use of the <om:result> element and SWE coding of an observation (upper branch) or link to a data viewer (lower branch).

data. Such a document should be divided into three sections: i) Characteristics, ii) Capabilities, and iii) History. Domain-specific parameters, such as depth of sampling, should be included within the first two, while the last one has great potential to let users understand the context in which the data has been acquired—for example, to store information on anomalies or events that occurred during acquisition. This section has already been exploited for the integration of this metadata model with the FP7 EU Eurofleets projects where the output of software devoted to automatic and manual event logging of surveys has been linked to this SensorML slot (Diviacco *et al.*, 2015).

Back to O&M, the <om:result> element is a preset that targets the observation itself. Similar to the strategy employed in the case of the <om:procedure>, the <om:result> element can also contain SensorML code holding values (Fig. 2, upper branch) or a URL to a downloading or preview service (Fig. 2, lower branch).

Where "om:result" holds SWE carrying a value, the corresponding XML code can be as follows:

```
<om:result>
    <swe:DataArray>
        <swe:elementType name="values">
            <swe:DataRecord>
                <swe:field name="Salinity_of_the_water_sample">
                    <swe:Quantity
                        definition="http://vocab.nerc.ac.uk/collection
                        /P01/current/PSLTZZ01/">
                        <swe:uom xlink:href="psu" code="psu"/>
                        <swe:value>31.77080</swe:value>
                    </swe:Quantity>
                </swe:field>
            </swe:DataRecord>
        </swe:elementType>
    </swe:DataArray>
</om:result>
```

When the <om:result> element points to a URL, this can be written as follows:

```
<swe:values xlink:href="http://www.utm.csic.es/SadoWS/DataQuery?
BaseDatos=SADO_SDG_RT&grupo=termosal&servicio=ultimo"/>
```

In the case that linked records hold multiple values, it is necessary to define the encoding of field identifiers, such as in the following example:

```
<swe:encoding>
        <swe:TextEncoding blockSeparator="&#13;&#10;" tokenSeparator=","/>
</swe:encoding>
```

The order of the values in the record will mirror the order of fields in the <swe:DataRecord> element:

```
<swe:DataRecord>
                <swe:field name="Salinity_of_the_water_sample">
                        <swe:Quantity
                        definition="http://vocab.nerc.ac.uk/collection/P01/current/PSLTZZ01/">
                                <swe:uom xlink:href="psu" code="psu"/>
                        </swe:Quantity>
                </swe:field>
                <swe:field name="Temperature_TS90_of_the_water_sample">
                        <swe:Quantity
                        definition="http://vocab.nerc.ac.uk/collection/P01/current/TEMPSD01/">
                                <swe:uom xlink:href="http://vocab.nerc.ac.uk/collection/
                                P06/current/UPAA"
                                code="Cel"/>
                        </swe:Quantity>
                </swe:field>
                <swe:field name="Raw_fluorometer_output">
                        <swe:Quantity
                        definition="http://vocab.nerc.ac.uk/collection/P01/current/FVLTZZ01">
                                <swe:uom xlink:href="http://vocab.nerc.ac.uk/collection/
                                P06/current/UVLT"
                                code="volts"/>
                        </swe:Quantity>
                </swe:field>
</swe:DataRecord>
```

6. The issue of the Feature of Interest

One of the most useful elements in O&M is the <om:featureOfInterest> (FOI), which carries the property which is observed [OGC_2]. Since a feature is a representation of a real-world object, each domain will shape the FOI in a different way, which of course will heavily condition data selection.

In order to enable multiple domains to refer to the same data set, even having started from different representations, we need a standardized but flexible means to use FOIs.

Contemporary sociology of science can provide a perspective in this through the ideas introduced by Star and Griesemer (1989). They suggested that when multiple communities, and therefore cultures, need to interact, notwithstanding the fact that they could have rather different cognitive models, they can effectively collaborate. In this, communities need to be gathered through the means of artifacts that Star and Griesemer call "boundary objects". These aim to bridge concurrent cognitive models through abstraction from all the domains of the partners. Boundary objects are then weakly structured in common use, while strongly structured in individual use. They contain sufficient detail to be understood by one partner, although it is not necessary that all partners understand the context in which the other partners use a boundary object. The perspective of the boundary objects has already been applied in the world of data management with a specific focus on collaborative work (Diviacco *et al*., 2012, 2015; Diviacco and Busato, 2013). We contend that the same perspective can be used here, applying it to the role of the <om:featureOfInterest> element.

In fact, defining multiple, generic, and subjective FOIs will allow multiple domain-specific and end user-oriented representation of the same portion of reality and observation. An example of this would be the water column and the seafloor structure in the seismic signals.

In the same way, a FOI can have different definitions, all of them valid for a specific data set. Using the same example of the water column, the FOI could be something generic like "water column", something more concrete like a collection of points "latitude, longitude, depth", or a general descriptive location, such as the Atlantic Ocean.

As a matter of fact, in the OGC SWE Sensor Observation Service (SOS) method GetCapabilities, the client starts by making the GetCapabilities request, which returns the Capabilities document of the SOS. By parsing and analyzing that document, the client gets references to related Features, along with other descriptions of content and services.

As in the case of search engines, where multiple keywords are used to better drive information discovery, the possibility of using multiple FOIs has been introduced to allow an easier definition and discovery of the data.

It could be a good approach to use the data user experience to enrich metadata. We call this practice dynamic metadata handling, as in the case of an end-user learning that the data is clipped – so he/she can add different FOIs to it.

To describe the context of data, other means can be used. SensorML standard defines conceptual models and an XML implementation of these models for describing non-physical and physical processes surrounding the act of measurement and subsequent processing of observations. The conceptual models are described using UML, while the implementation is described using the XML Schema language and Schematron.

To share terminology, it is important to use a controlled vocabulary. In the SeaDataNet and GeoSeas case, this is guaranteed by a web link to the corresponding term in the BODC vocabulary:

```
<swe:field name="Salinity_of_the_water_sample">
    <swe:Quantity definition="http://vocab.nerc.ac.uk/collection/P01/current/PSLTZZ01/">
        <swe:uom xlink:href="psu" code="psu"/>
    </swe:Quantity>
</swe:field>
<swe:field name="Temperature_TS90_of_the_water_sample">
    <swe:Quantity definition="http://vocab.nerc.ac.uk/collection/P01/current/TEMPSD01/">
        <swe:uom xlink:href="http://vocab.nerc.ac.uk/collection/P06/current/
        UPAA" code="Cel"/>
    </swe:Quantity>
</swe:field>
```

7. Conclusions

This work details how the problem of mis-linking discovery and usage of data due to contrasting cognitive models can be addressed. We contend that the solution can be found within an approach that integrates data discovery, browsing, and access through machine-harvestable, content-rich metadata. To allow this, the designated infrastructure should rely on a layered metadata model that can detach but at the same time link generic data discovery and domain-specific search paths, while allowing data preview. We implemented this approach in the framework of the SeaDataNet Project, extending and adapting it—after it had been successfully tested in the domain of geophysics—to oceanographic ideas.

REFERENCES

Baumann P.; 2014: *No more metadata!* In: EGU conference 2014, <http://presentations.copernicus.org/EGU2014-15508_presentation.pdf>.

Diviacco P. and Busato A.; 2013: *The Geo-Seas seismic data viewer: a tool to facilitate the control of data access*. Boll. Geof. Teor. Appl., **54**, 257-270.

Diviacco P., Lowry R. and Schaap D.; 2012: *Marine seismic metadata for an integrated European scale data infrastructure: the FP7 Geo-Seas project*. Boll. Geof. Teor. Appl., **53**, 243-252.

Diviacco P., De Cauwer K., Leadbetter A., Sorribas J., Stojanov Y., Busato A. and Cova A.; 2015: *Bridging semantically different paradigms in the field of marine acquisition event logging*. Earth Science Informatics, **8**, 135-146.

Forbriger T.; 2003: *Inversion of shallow-seismic wavefields:I. Wavefield transformation*. Geophys. J. Int., **153**, 719–734.

Hamon M., Reverdin G. and Le Traon P.-Y.; 2011: *Empirical correction of XBT data*. <http://www.coriolis.eu.org/content/download/9468/63589/file/XBTos--hamon.gr.V1.05082011.pdf>.

Schaap D., Lowry R.K.; 2010: *SeaDataNet - Pan-European infrastructure for marine and ocean data management: unified access to distributed data sets*. International Journal of Digital Earth, **3**, 50-69, doi: 10.1080/17538941003660974.

Star S. L. and Griesemer J.R.; 1989: *Institutional ecology, translations and boundary objects: amateurs and professionals in Berkeley s museum of vertebrate zoology*. Social Studies of Science, **19**, 387–420, doi:10.1177/030631289019003001.

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/

[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
[O&M] http://www.opengeospatial.org/standards/om
[OGC] http://www.opengeospatial.org/
[OGC_2] http://www.ogcnetwork.net/sos_2_0/tutorial/om
[P1/90] http://www.epsg.org/exchange/p1.pdf
[SeaDataNet] http://www.seadatanet.org/

Corresponding author: Paolo Diviacco
Istituto nazionale di Oceanografia e di Geofisica Sperimentale (OGS),
Borgo Grotta Gigante 42c, 34010 Sgonico (TS), Italy
Phone: +39 040 2140380; fax: +39 040 327307; e-mail: pdiviacco@inogs.it