Bridging the gap between data and metadata (Part 1): a novel approach in Metadata, Data models and Semantics emerging from addressing the issues of seismic data within the Geo-Seas project.

Paolo Diviacco*
Alessandro Busato
Helen M. Glaves
Dick M.A. Schaap

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Borgo Grotta Gigante 42/C, 34100, Trieste, Italy (pdiviacco@ogs.trieste.it)
Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Borgo Grotta Gigante 42/C, 34100, Trieste, Italy (abusato@ogs.trieste.it)
National Environment research Council (NERC) Nottingham, UK (hmg@bgs.ac.uk)
Mariene Informatie Service 'Maris' BV The Hague Area, Netherlands (dick@maris.nl)

1. The EU Fp7 Geo-Seas project

The European Union 7th framework project Geo-Seas aims at building an European data space in the field of marine geophysics and geology. This data space is to be integrated with the existing SeaDataNet oceanographic initiative 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, OneGeology and Eurofleets, into account and is commited to refer to and stimulate the EU Inspire directive through its developments

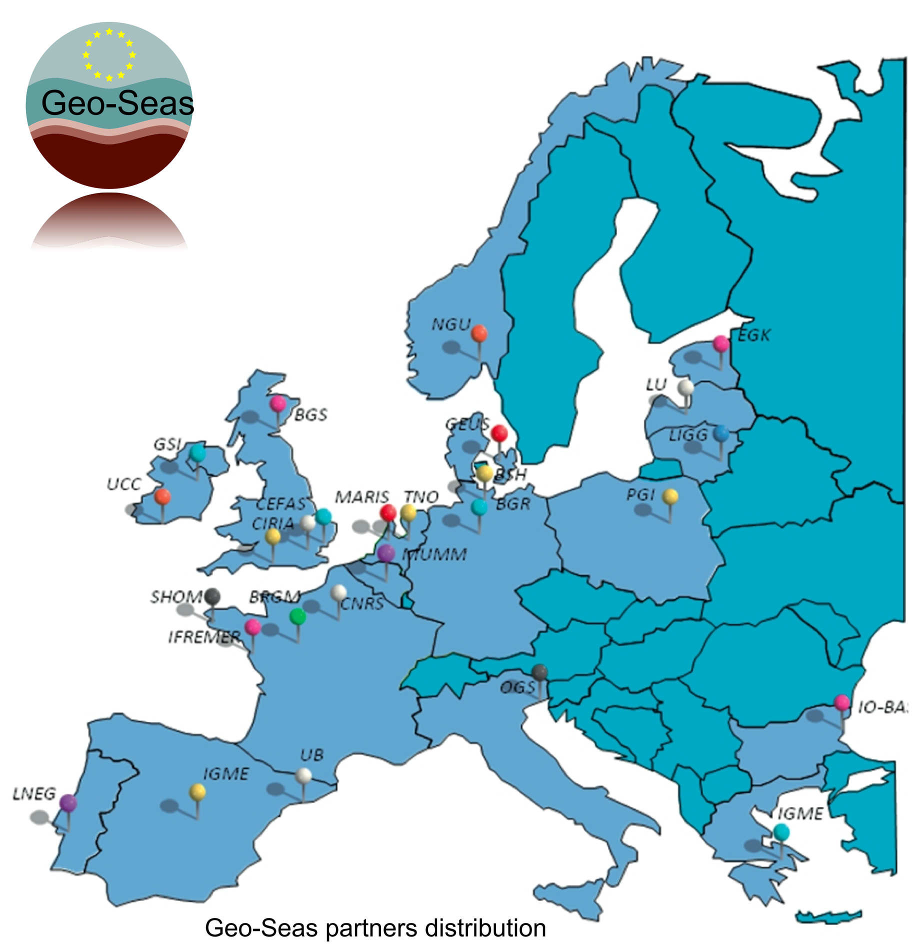
Examples of data types that can be delivered by Geo-Seas are seismics, bathymetry (including digital terrain models), lithology, mineralogy, geochemistry, sediment grain-size and geotechnical data.

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.

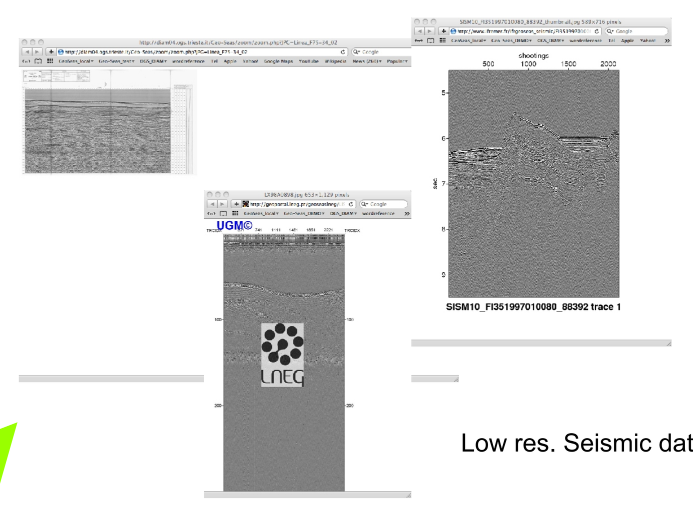
The user experience in such a data space spans all steps from data discovery to data access and is supported by a mosaic of metadata models, data formats, vocabularies and services.

Despite well-established cooperation between national geological surveys and research institutes active in the European seas, so far, it was not possible to federate separate national databases since these were built using different nomenclatures, reference levels, formats, scales, coordinate systems and practices in general.

This hampers direct integrated use of 'primary' data and the generation and dissemination of trans-boundary or multi-disciplinary datasets, products and services.

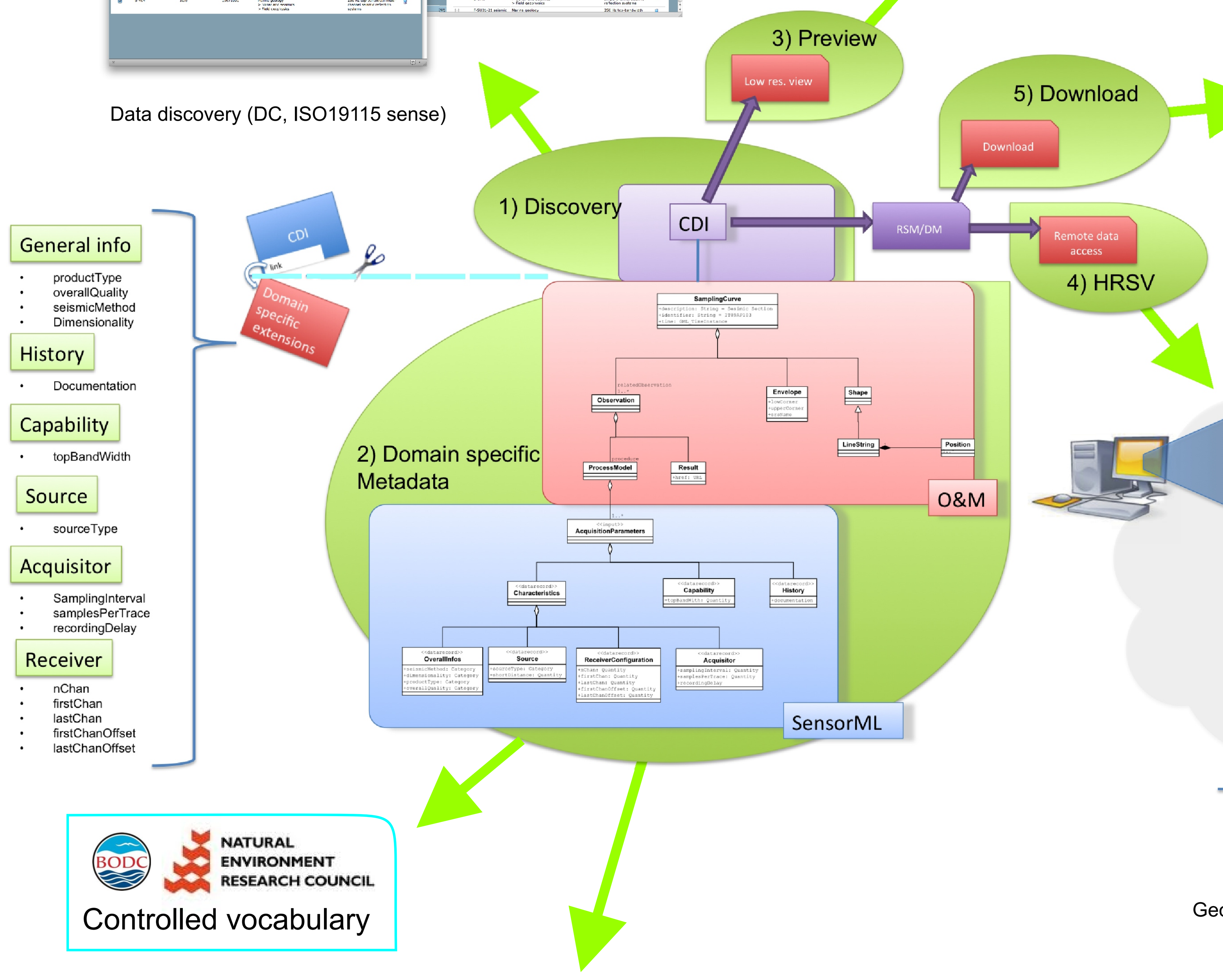


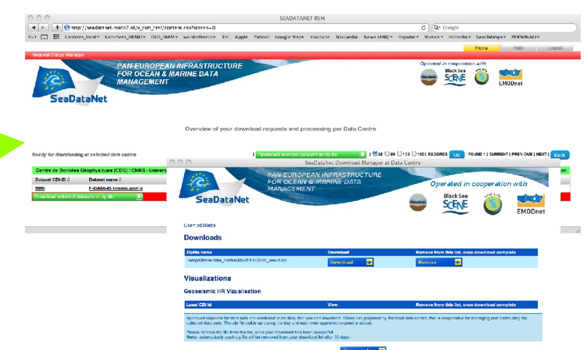
Geo-Seas partners distribution



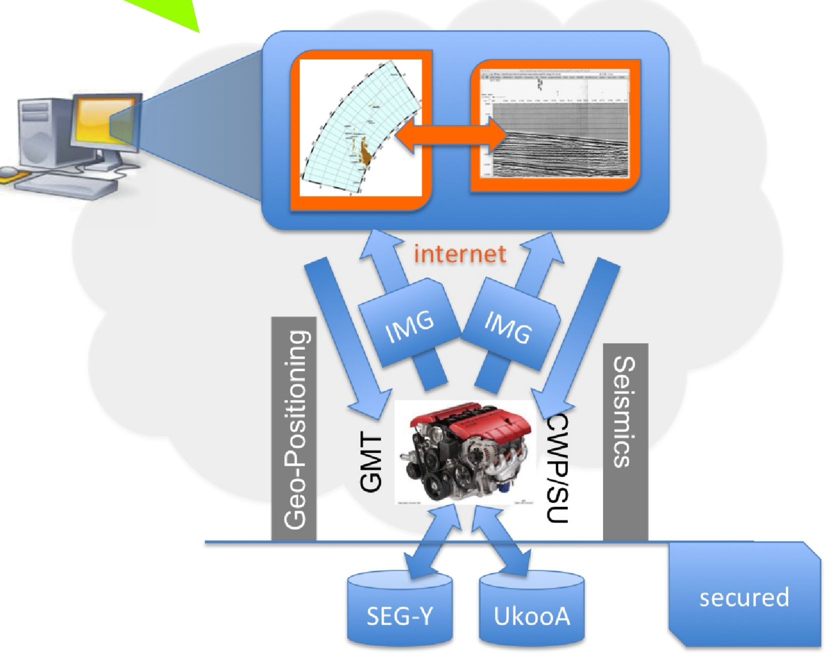
Low res. Seismic data preview examples

Data discovery (DC, ISO19115 sense)

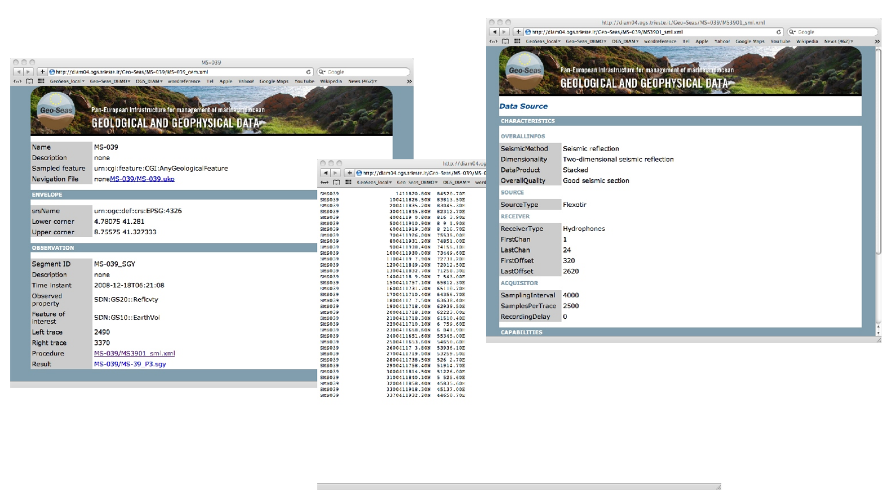




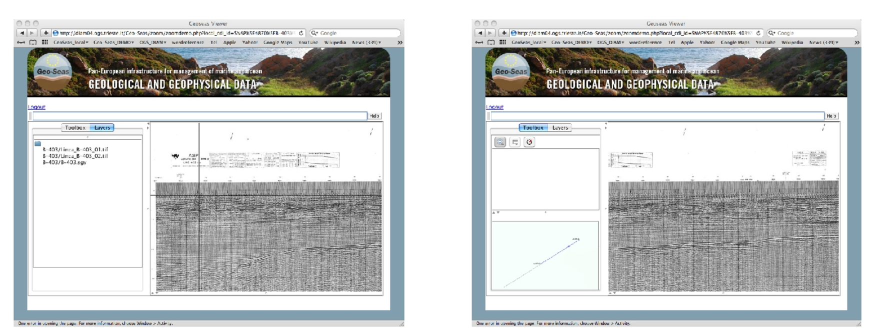
Geo-Seas downloading facitity



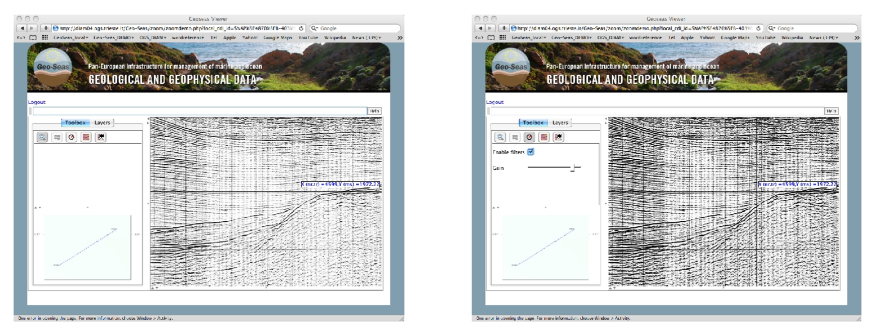
Geo-Seas HR seismic viewer software structure



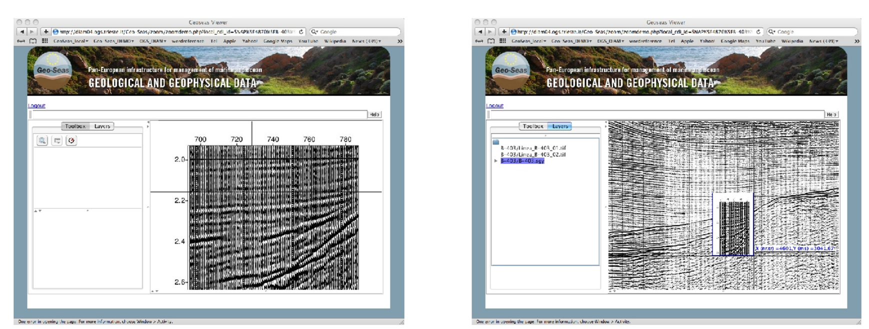
Domain specific metadata, positioning and O&M, SensorML transformed by XSLT



Geo-Seas HR seismic viewer tiff and Seg-Y visualization



Geo-Seas HR seismic viewer Gain processing



Geo-Seas HR seismic viewer wiggle plotting and superimposition

2. Seismic data

In comparison with other data types handled by the Geo-Seas initiative, seismic data has particularities that need to be considered carefully.

From a technical point of view most of the issues are related to the dimension of the data files. In fact, although the extent of a seismic suvey can vary a lot from case to case, file sizes of field data very easilly can exceed the GigaByte. It is therefore easy to immagine that moving through the web such amount of data has to be considered carefully.

Another very important point of view is that of the data value. In fact seismic data acquired by research institutes, being very important in oil and gas exploration can be an important resource for oil companies and therefore has to be handled with care to avoid that the data owner loose the control on the data itself. In fact once downloaded there's no way of protecting data.

At the same time data is one of the most important facilitators in fostering the collaborative attitude of research institutions. It is therefore very important to balance protection and dissemination of data.

3. Finding the right data

Information overload is and important issue to consider in any research, and in the field of geophysics it could be even more relevant considering the costs of preparing the work of scientists, processing and loading data into vertical interpretation software. To avoid this, careful selection of data is a mandatory while absolutely not trivial task.

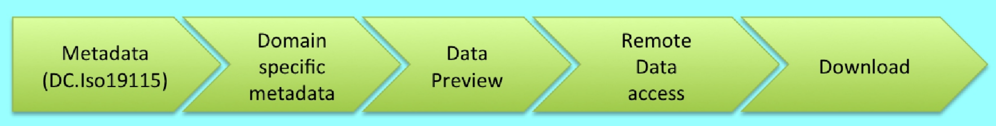
Data discovery has been traditionally separated from data usage. Metadata is a description of data based on semantic rich text supposed to be queried in order to filter all available datasets to hit the data an end user is interested in.

On the opposite side data is simply the observation itself, without any semantics.

Traditional discovery of data is intended in a Dublin Core or ISO19115 sense; meaning a core set of metadata elements that allows searching across domains or data types.

In our opinion this duality between metadata and the actual data is detrimental to an effective data search and therefore we propose to bridge these two extremes.

In this perspective within the GeoSeas project we identified 5 steps that allow an end user to be fully aware of the usefulness of the data he/she hit.



Five steps in bridging metadata and data

These steps are:

1) Core metadata (DC, ISO19115 sense): cross domain minimal information as when, where, who, simple informations on what. This can correspond to a single profile crossing all domains

2) Domain specific metadata: domain specific parameters (e.g. Sampling frequency in the case of seismics). It has to be noted that considering that the set of metadata changes from domain to domain, it is not possible to have just one cross-domain profile

3) Data preview: once possible data have been identified upon core and domain specific metadata, it is necessary to step into data itself, but considering the above mentioned technical and value-related problems of seismic data this has to be done through several steps. The first of which is data preview where data is represented as a low resolution and possibly watermarked static image of the data. Here large geological features can be seen but not small ones.

4) Remote data access (viewer): if data preview revealed that the data could be of interest then the end user needs to have a look at subtle feature and this cannot be done using static images but only with a data viewer that should on the other hand allow protection of the data.

5) Download: once data has been identified, the end user should contact the data owner and negotiate the possible use of the data. Generally this corresponds to an agreement that include possibly a joint collaborative project on the exploitation of the data.

4. Implementation

The SeaDataNet infrastructure had already made available tools and services to allow discovery (1) and downloading (5) of data, the Geo-Seas project has had to develop a further three segments and integrate them in the infrastructure for use with seismic data.

Browsing was addressed introducing O&M and SensorML, with a special care to the definition of a domain specific metadata profile and related controlled vocabulary (Diviacco et al. 2012) that, linked from the CDI, allows detailed selection of data. Data preview was implemented as static images, while a lot of efforts were spent developing the web based tool aiming at allowing direct interactive access to the data (Diviacco2005, Diviacco & Busato 2013).

This is based on a mixed server side client side paradigm where interactive visualization of seismic data is georeferenced to SVG based geographic maps.

Conclusions

The Geo-Seas data space adopted a five step data discovery and access procedure that can effectively bridge the gap between metadata and data access, avoiding information overload and allowing end user to rapidly hit the data they are interested in.

The introduction of a High resolution Viewer allowed to balance at the same time data protection and dissemination.

We think that this perspective although born to address the above mentioned specific technical and value-related issues of seismic data can be usefully extended in other domains and to other data types.

REFERENCES

Diviacco, P., 2005 "An open source, web based, simple solution for seismic data dissemination and collaborative research", Computers and Geosciences, 2005, Vol 31/5 pp 599-605

Diviacco, P. 2006 "Integrating Scalable Vector Graphics (SVG) In a web based seismic data Portal", EAGE 2006, Vienna

Diviacco, P. et al.¤2012 Marine Seismic Metadata for an Integrated European Scale Data Infrastructure. The Fp7 Geo-Seas Project,¤BGTA, DOI: 10.4430/bgta0051

Diviacco P., Busato A., "The Geo-Seas Seismic data viewer: a Tool to facilitate the control of data access" BGTA vol.54/2 june 2013

[1] SVG: http://www.w3.org/Graphics/SVG

[2] Batik: http://xmlgraphics.apache.org/batik

[3] CWP/SU: http://www.cwp.mines.edu/cwpcodes

[4] JMS: http://docs.oracle.com/javaee/1.3/jms/tutorial