Mediterranean Marine Science
Indexed in WoS (Web of Science, ISI Thomson) and SCOPUS
The journal is available on line at http://www.medit-mar-sc.net

Paper presented at the
"International Conference on Marine Data and
Information Systems - IMDIS", Athens, 2008

Towards a collaborative research data space in Geophysics

P. DIVIACCO

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS),
Borgo grotta Gigante 42/c, 34011, Trieste, Italy

Corresponding author: pdiviacco@ogs.trieste.it

Abstract

Data management means managing data as a valuable resource. Unused data have no value and data value depends on how they are used. This extends the focus from the data itself towards the processes to which they belong. Within this perspective, we will here focus on scientific research in Geophysics, albeit probably the discussion can be generalized to other fields of Science.

Modern epistemology describes Science as a social construct. In this it is important to understand the role data have in scientists life but also the human and social aspects researchers project on data. Scientists need means to support their theories but at the same time data owners need to preserve the investments necessary to acquire data. The data systems available by now, generally, do not satisfy the needs of the latter so that these are reluctant to share their assets. Novel perspectives and tools are needed that allow data owners to balance protection and the need to open their archives to attract new collaborations. This is possible within a mixed server side-client side paradigm where nonsensitive data are sent to the client while data to be protected can be accessed through the web directly using the data owner IT system without moving or copying files to the user computer. We will here report on how this view can be translated into a working system.

Keywords: Data management; Geophysics; Data protection; Scientific collaboration.

Introduction

Researchers are individuals rooted in different disciplines, traditions, schools, cultures and some may say also generations. Sometimes, their myths and models about scientific research do not correspond to their "normal" scientific life, plus, some of them are not even aware that this could happen.

Modern epistemology undermined the theoretical framework of the traditional view of Science. The new view is shaped on a line of thought that connects authors as Duhem, Lakatos, Latour and others. Science is not a "cold", mechanical, "undebatable" process, instead, following this view, it is profoundly conditioned by human factors, and since researchers are part of communities, eventually, as in Latour (LATOUR, 1987), Science is a social construct.

Following Kuhn (KUHN, 1962), scientists live within paradigms, a philosophical or theoretical framework, a tradition or school, that condition their way of thinking. Different, concurrent and incommensurable paradigms exist within any discipline.

The GeoSciences fit perfectly with this vision. As in Frodeman (FRODEMAN, 1995) they are based on the observation and interpretation of limited datasets resulting from complex "natural experiments" which very unlikely can be repeated in a controlled laboratory work. This allows concepts to be constructed that are then used to interpret data. The concepts are theories or models that often become accepted dogmas for certain regions or settings. In the case of Seismic data, following Bond (BOND *et al.*, 2007), interpretation is based on previous experience and preconceived notions that stem from the personal background of the scientist.

Following Becher (BECHER *et al.*, 2001) scientists gather in communities that resemble "tribes". Here social and institutional characteristics of knowledge communities matter for the epistemological properties of the knowledge they produce (their territories). Conditions and circumstances that are external, like how universities, faculties and departments are organized, or as in Whitley (WHITLEY, 2000) the control over facilities, can and do make a difference to disciplinary status and identity.

Within this context, the control over Data, which are the primary sources of research, is of paramount importance. How data are used is reflected in any following secondary source of research as publications, but also in practices, trends, reputation and ultimately it determines also how funding are distributed.

Up to now primary and secondary sources of research tended to be decoupled in the life of scientists. Observations remained in the basement archive, when not in a drawer, and theories were discussed in meeting or journals, showing edited results and without the possibility to go back to the original observations. Most of the existing data systems only improve the efficiency of this traditional view, while new technologies offer means to introduce an innovative perspective where both activities can be captured in the same space. Using hypermedia, researchers will crosscorrelate simultaneously theoretical trains of thought with data, experiencing a novel way of making collaborative science.

The role of data

Only very seldom researchers have at their disposal all the data they need. Acquisition is very expensive, especially within the field of geophysics, so that surveying should be minimized reusing data acquired by other institutions (MILES *et al.*,2007). Often researchers, find particularly interesting the most extreme environments where acquisition can be highly uncomfortable when not impossible at all, so that surveying a specific area, sometimes, can be done only once. Moreover to study the evolution of a phenomenon, timelapse or historical recording should be compared. All this fosters the practice of networking collaborative work among scientists and institutions which is warmly welcome and deeply encouraged by funding agencies all over the world.

In this, two kind of actors emerge: Data owners and Data seekers.

The latter "simply" needs data and the tools to find them, while the role of the former is more problematic. Data owners invested money, time and resources in the acquisition of their datasets and therefore, considering all we said above, it is natural that they are generally reluctant to lose the control on Data and wish to take part of the payback, as for example publications.

At the same time they need to open their archives to trigger new collaborations, to publish papers, and eventually to attract new funding. They need therefore a solution that allow them to balance protection and dissemination.

Many initiatives exist in the field of exchange and dissemination of Geophysical data. Most of them focus on the needs of data seekers, while providing no mitigation for the anxieties of data owners. This eventually results in the latter submitting as few data as possible when not trying to avoid the issue at all. Contractual obligations can be easily bypassed since it is very difficult to impose to data owners the quality and format they should adopt. For example, quite often instead of the actual data, the owner submit low resolution images, or reports about the data.

A tempting solution can be to share metadata only, leaving data to further direct negotiation once the actors were put in touch.

This may work for some data type and disciplines, but for Geophysics it is not efficient and results in a huge amount of time wasted. In fact scientists need to assess the value of the data they are considering before drawing up agreements with the data owner. This value depends on many variables. Generally researchers have a clear idea of what they are looking for and since there is no simple way of capturing such features in the metadata, they need to visualize data.

In this perspective, data quality is decoupled with data value since the former has nothing to share with the presence or not of a feature a scientist is looking for. Often, also very noisy data can be useful, plus sometimes noise itself can be a mean of detection, as in the case of modern treatment of seafloor multiples, surface waves inversion or diffraction migration. Again, this suggests that visual inspection of data is the only useful way to assess the usability of data.

On the other hand the need to view data is not easy to be fulfilled in Geophysics. Here in fact, and particularly in the case of Exploration Seismics, data correspond to large files, where it is common to exceed the Gigabyte. Considering that, nowadays, users are accustomed to systems that react very quickly, this, of course, advises against the network transfer of data outside the scope of a Local Area Network, while of course any international collaboration assumes that scientists can be distributed all around the world.

This, which seems at first sight a big limit, actually can trigger a novel perspective that could help solving the above mentioned data owners puzzle of balancing protection and dissemination.

The method: a mixed server-side client-side paradigm

The distinction between data owners and seekers and their characteristics can be translated into the relationship between their computers. One is offering something the other is asking something. This resembles a client/server architecture, where data resides on a system managed by the data owner, and the client is any computer owned by the data seeker. Of course they should be connected through the Web in order that virtually there should be no geographic limitation in setting up the collaboration. The problem here is what data can be transmitted, from the server to the client without loosing control on them and at the same time which computer will carry out the processing.

In this, it is of help to identify the activities. In the case of Geophysics, it is useful to class them on the request of a geographic (mapping) or of a depth/vertical distribution (visualization). Especially in the case of Exploration Seismics, this is of help because the actual data, being depth (or TWT) sections, express the Z axis while the metadata (mainly positioning) are to be mapped geographically so that it is easy to separate data from metadata and handle them differently.

At least in the Scientific Research Environment, data position is not a sensitive issue, and therefore can be sent to the client. Though, this is not the common practice today. Most of the software solutions for web based mapping, in fact, are built on a server side paradigm. Maps are compiled and rendered in the server and only then sent to the web browser very much like a standard html static image with hot spots. Each time the user wants to zoom or pan he/she sends a request to the server that again compiles and renders a new image that is sent back to the user.

This, besides increasing the network traffic, overloads the server.

A different approach can be implemented moving the rendering activity to the client computer. This can be achieved provided that the client has a mapping tool.

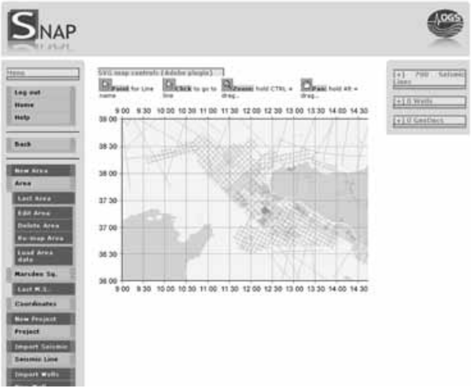

Fig. 1: A web page embedding an SVG map showing the position of several seismic lines acquired in the area

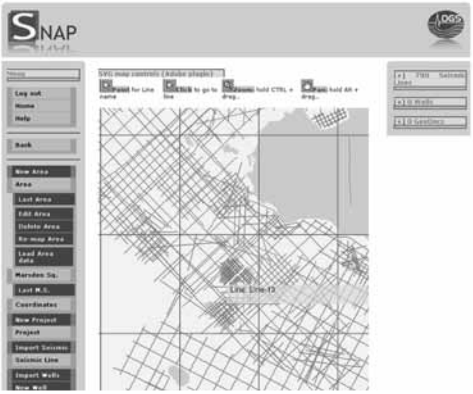

Fig. 3: Within an SVG map objects remain such. They are isolated and react independently. Here in a quite cluttered area it is easy to identify a seismic line. Clicking on it the system sends the user to a page listing the data available.

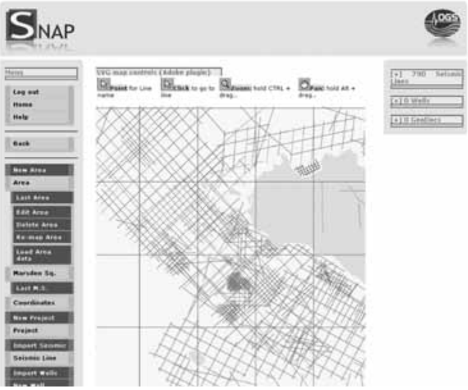

Fig. 2: Zooming in the SVG map there is no loss of resolution

Nowadays this is quite not a problem. Many options exists from Virtual globes as Word Wind or Google Earth to GIS software using WMS/WFS/WCS/GML.

To avoid confusing users with additional software (and possible license), we decided to develop a solution to be selfcontained in a web browser (Fig. 1).

This solution is based on Scalable Vector Graphics (SVG), an open standard created by the W3C's SVG Working Group, which is an XML specification and file format for describing twodimensional vector graphics that may include scripting.

There are many advantages of an SVG based mapping tool compared to other solutions. SVG is a vectorial format meaning that while zooming, the viewer renders objects without loosing resolution (pixellation) (Fig. 2).

Within an SVG map, objects remain such, meaning that they are not only a bunch of pixel but preserve their identity, so that they can be separated even when crossing other objects. This way they can be associated with specific attributes, for example a web link. Objects, as for example seismic lines, can be have methods, so that, for example, to easily highlight their areal extent, hovering on them, these can change color and a box with the object name can pop up (Fig. 3).

Another great advantage of SVG is that since the rendering is carried out by the user computer it can take advantage of the resources of this latter. Since any commodity computer nowadays mounts powerful video cards, mapping becomes very fast while at the same time unloading the server and the network.

Of course a first phase of map compilation and SVG file creation has to take place serverside, however also this lag can be reduced using "smart" serverside caching systems. After some

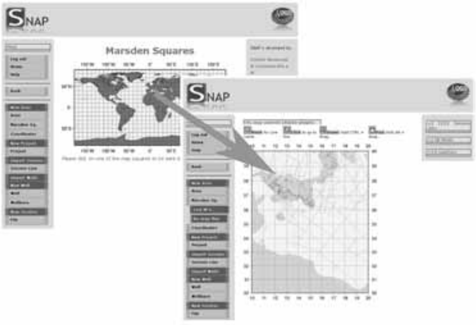

Fig. 4: Using Marsden Squares it is possible to precompute maps to be cached server-side in order to reduce the time required for compilation and rendering of a map created from scratch.

usage monitoring, in fact, we realized that users do not query at random. Instead, they tend to focus on a quite narrow subset of all the possible queries. These of course varies depending on the database content, but very likely depend also on the current stream of interests within the institution that manages the system. This allows to prepare a certain amount of precomputed maps, that can be suggested in order to avoid mapping from scratch. Our experience confirms that generally this option is not perceived as a limitation but on contrary as a useful way of saving time.

In the case of Exploration Seismics targeted to large geological structures we found very useful the introduction of Marsden Squares (Fig. 4), a mapping system, used by the World Meteorological Organization (WMO), based on grid cells of 10° latitude by 10° longitude, each with a unique, numeric identifier, while for studies regarding a specific geologic province, it is more useful to define exactly its areal extension. Once the areas are defined, the data manager can produce a precomputed map that will automatically be updated when new data is uploaded. These Provinces or Marsden Squares are listed upon login and can be easily accessed by any user.

A further level of "smart" server side caching can be implemented to recycle previous requests. We realized that often requests made by a member of a workgroup are requested again later by other members of the same group or even by other users, so that if previous processing is cached (server-side) it can be sent again without loosing time creating the file once more.

When a seismic line is selected we start facing the problem of balancing protection and dissemination. Most of the solutions available by now are based on a downloading practice. Here users connects to a repository where files can be copied to their computer. In many cases these repositories are managed by an international initiative where data owners are supposed to send their data. It is easy to understand that this way there is no possibility for data owners to protect their assets, nor physically, neither legally.

A different perspective is needed, which avoid the downloading practice while keeping data server side all the way through the access to them. This can be done if the data access facility offers all the means to use the data in a protective environment, where users can access the data they are entitled to, after agreement with the data owner. Users that did not received a formal permission can access only low resolution and/or watermarked images of data, so that these cannot be used in any publication or scientific work.

Full access to the data can span from simple services where data are only visualized, to more complex ones where data can also be processed, until the most sophisticated where data can be interpreted "on line". In this direction a new class of web based collaborative software will be able to support all the relevant activities related to collaborations among international partners.

Diviacco (DIVIACCO, 2005) proposes a simple system that can address the needs of visualization of digital seismic data recorded using the SEGY format (BARRY *et al.*,1980), an open standard developed by the Society of Exploration Geophysicists and widely used within the geophysics community. The visualization system uses common web based tools and open source vertical applications that can be orchestrated using PHP. This system relies on MySQL to store on one side all the relevant metadata and

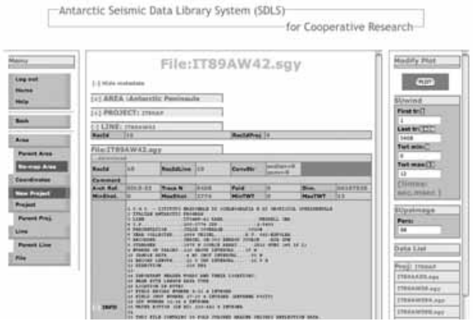

Fig. 5: Metadata page related to a seismic line.

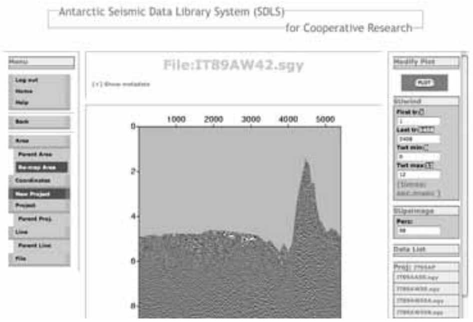

Fig. 6: Seismic data visualization.

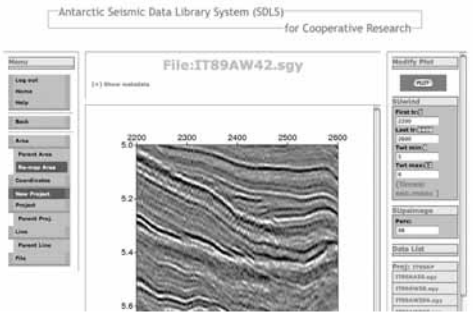

Fig. 7: Seismic data can be zoomed and panned interactively to reveal any useful data feature.

on the other side the state of the system itself (Figs. 5, 6, 7).

This framework can be extended to handle other types of geophysical data as wells or "vintage" seismic sections. In both cases this generally means scanning old paper plots and related documents (Fig. 8).

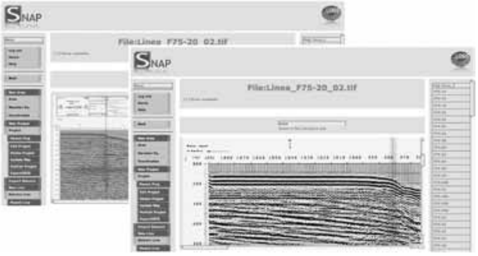

Fig. 8: Visualization of "vintage" scanned seismic sections.

As in the Seiscan (MILES *et al.*,2007) experience the tiff file format can be used to balance quality and file size, but since this format is not compatible with any standard web browser it is useful to convert these files to png, gif or jpg to allow a quick lookup visualization. This can be done automatically when uploading files to the system. Within the same perspective it is possible to handle also any kind of document related to a geographic feature, as for example a well log (Fig. 9), mapping it to allow users to retrieve it with a simple click.

Systems based on such mixed server side-client side can be deployed at or developed by any existing initiative aimed at disseminating geophysical data.

OGS is currently offering this service within some international scientific data management initiative as for example the Antarctic Seismic Data Library System (SDLS) or the European Consortium for Ocean research Drilling (ECORDnet). Our experience confirms that allowing owners to control the usage of their data foster their attitude to share their assets with the designated community.

We are currently exploring a different approach where standalone systems can be installed at any institution requiring it, allowing them to independently manage their collaborative data space. To organize such a decentered system and allow any user to search within it, a central harvesting facility is needed, able to

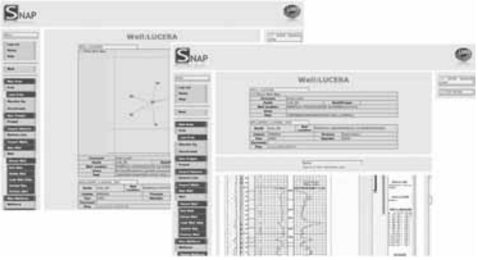

Fig. 9: Visualization of well logs.

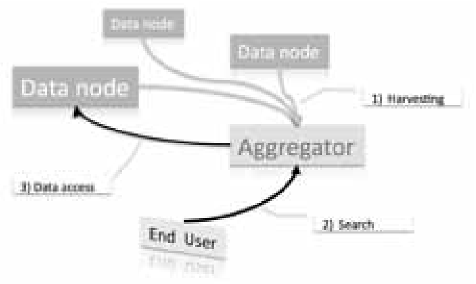

Fig. 10: A distributed system where users can query an aggregator that sends them to the remote node offering access to the actual data.

collect automatically positioning and metadata from federated systems, publishing them as above but redirecting data seekers to the satellite system for data access (Fig. 10).

Metadata

Scientific Research in Geophysics and the E&P industry have always been crossfertilizing contexts, the differences between them emerging more towards the tail of the exploration process. It was natural then to adopt common data standards as SEGY (BARRY *et al.*, 1980) for the actual seismic recordings and UKOOA (UKOOA exploration committee, 1991) for positioning.

These formats not only contain the actual data but also most of the relevant metadata, or at least what is really needed to use them. Addressing most of the practical issues of geophysical exploration, in spite of being quite old, these standards are still widely used within the community, which has the side effect that pure metadata is not perceived by scientists as absolutely necessary. Here the E&P industry and Scientific research diverge. Important initiatives exist as the Public Petroleum Data Model (PPDM) or the Petrotechnical Open Standard Consortium (POSC) which are mainly focused on the industrial value of data. If these offer a comprehensive and complex structure that perfectly fits the needs of the E&P industry (upstream and downstream) they are simply too much for Scientific Research. Conforming data to these international standards means, in fact, appointing resources and personnel which requires investments that are far beyond the possibilities of most of the research institutions involved in this field. Here data management is largely based on the voluntary work of some single "enlighten" researcher, which often lacks the time, the tools and sometimes the knowledge to correctly drive through all the requirements of international data and metadata standards.

Much more lightweight metadata models are therefore needed. These should conform to the requests of Core ISO19115, with extensions to report basic geophysical metadata. Unfortunately up to now no standardization effort in this sense was widely accepted.

Considering the above mentioned problem of resource allocation for data management experienced by research institutions and the widely acceptance of SGY/UKOOA metadata model, we decided to use Core ISO19115 while at the same time extract metadata (as for example sampling rate or fold coverage) from SGY (BARRY *et al.*,1980) and positioning from UKOOA (UKOOA exploration committee, 1991) or SGY (BARRY *et al.*,1980) itself. Metadata extraction can be performed automatically when data is uploaded without the need of any human interaction

Conclusions

Data control is a powerful mean to position any institution within a scientific community. Up to now web based data systems and management initiatives in the field of Geophysics do not offer data owners means to address such an important issue. This results in a retentive policy that leaves great amounts of valuable data outside the shared data space of the designated community.

A novel view was here presented that based on a mixed server side-client side paradigm can mitigate the anxieties of data owners and at the same time is able to fully exploit all the technological aspects involved in web based Geophysical data access. The experience maturated within several international data management activities confirms that this view is well accepted by the actors and at least in the field of Geophysics, it can be of great help in establishing a shared and collaborative data space.

References

BARRY, K.M., CAVERS, D.A. & KNEALE, C.W., 1980. Recommended standards for digital tape formats, *Society of Exploration geophysicists*, Tulsa, Oklahoma.

BECHER, T. & TROWLER, P.R., 2001. *Academic Tribes and Territories, Intellectual Enquiry and the Culture of Disciplines*, Buckingham & Philadelphia: The Society for Research into Higher Education & Open University Press.

BOND, C.E., GIBBS, A.D., SHIPTON Z.K. & JONES, S., 2007. What do you think this is? 'Conceptual uncertainty' in geoscience interpretation, *GSA today*, 17: 4-10.

DIVIACCO, P., 2005. An Open Source, web based, simple solution for seismic data dissemination and collaborative research., *Computers & Geosciences,* Volume 31, Issue 5: 599-605.

FRODEMAN, R., 1995. Geologic Reasoning: geology as an interpretive and historical science, *Geological Society of America Bulletin*, 107: 960-968.

KUHN, T., 1962. *The structure of scientific revolutions*, University of Chicago Press, Chicago.

LATOUR, B., 1987. *Science in Action: How to Follow Scientists and Engineers Through Society,* Cambridge, MA, Harvard University Press.

MILES, P.R., SCHAMING M. & LOVERA, R., 2007. Resurrecting vintage paper seismic records. *Marine Geophysical Researches*, 28-4: 319-329.

UKOOA exploration Committee, 1991, UKOOA P1/90 post positioning data format, *First Break*, 9-4: 172-178

WHILTEY, R. 2000. *The intellectual and Social Organization of the Sciences*, Clarendon Press, Oxford.

SVG http://www.w3.org/Graphics/SVG/

PPDM http://www.ppdm.org/

POSC http://www.energistics.org/home