Oceanographic Data Repositories: An Analysis of the International Situation

Fabiano Couto Corrêa da Silva [1],*, Ernest Abadal [2] and Enrique Wulff [3]

[1] Institute of Humanities and Information, Federal University of Rio Grande (FURG), Av. Italia, Km 8, Rio Grande-RS 96200-400, Brazil

[2] Faculty of Library and Information Science, University of Barcelona (UB), Melcior de Palau, 140, Barcelona 08014, Spain; abadal@ub.edu

[3] Marine Sciences Institute of Andalusia, Spanish National Research Council (CSIC), Campus del Río San Pedro, Cadiz 11510, Spain; enrique.wulff@icman.csic.es

* Correspondence: fabianocc@gmail.com; Tel.: +55-53-999-776070

Abstract: The preservation and organization of oceanographic research data enables the scientific community to consult and reuse information of different kinds, and this is made possible by the repositories, meaning the services that facilitate data storage and dissemination. This paper reviews the current situation of oceanographic data repositories across different countries and evaluates them according to a series of indicators. The writers propose that although interest in storing and reusing oceanographic data has increased in recent years, the repositories are still in the process of developing their systems for processing, disseminating and reusing data. The repositories also differ in terms of architecture and the organizational level of the content they offer.

Keywords: data repositories; oceanography; evaluation; research data

1. Introduction

In their research, whether for academic, military or commercial purposes, oceanographers use a range of methods and instruments to generate large quantities of data on often dissimilar marine environments, from deep sea trenches to coastal regions, and from estuary systems to the Antarctic ice sheet.

In turn, oceanographic research embraces a wide range of disciplines, including the study of currents, waves and tides, the geological structure, composition and history of the ocean floor, and the chemical composition of bodies of water; its practitioners study mineral resources and marine biology, biodiversity and ecology. This variety is responsible for the diversity of the data that are stored and disseminated by the repositories in terms of spatial and temporal coverage and resolution, and the structure of the variables that repositories account for.

This new generation of high-performance scientific instruments, which includes telescopes, satellites, wireless sensor networks, accelerators, supercomputers and simulators, is also generating so much data that research communities now require an integrated e-infrastructure to globally generate and curate research data produced in very different conditions. Thanos [1] emphasizes the importance of integrating the development of the scientific community's data systems in Global Research Data Infrastructure (GRDI) and also describes a conceptual framework for GRDIs and a core set of functionalities that they should offer.

This need for integrating the e-infrastructure in which data repositories operate has been clearly demonstrated. The two-year EU project PARSE.Insight [2] highlighted our understanding of the major threats to the long-term preservation of research data, which can become inaccessible when hardware,

software or support systems are not sufficiently sustainable, when the provenance and authenticity of data is uncertain, when the location of the data cannot be identified or when it simply ceases to exist. For these and other reasons cited by PARSE.Insight, the digitization of research material whose data or data support systems are in some way imperiled must be one of the most important tasks in the preservation of oceanographic research data. In this way, it is necessary to mention Research Data Alliance (RDA) [3] in order to build the social and technical bridges that enable open sharing of data across technologies, disciplines and countries.

In order to exploit large quantities of data and make them useful, the scientific community requires a hierarchical structure in which collections are created within larger groups of similar collections and where data are properly registered and endowed with a descriptive body (metadata) that includes, amongst other elements, the legal authorization to access and disseminate data content.

Belter [4] stresses that data curation efforts need to be rewarded if we are to increase citations. According to his study, oceanographic datasets can obtain a higher number of citations than papers published on the same topic.

Metadata facilitates the discovery of data by including them in metadata repositories, such as the National Change Master Directory (GCMD) of the National Aeronautics and Space Administration (NASA). These repositories are essential for creating data packet registers, whereby any data can be traced back to their source.

The community is now collecting these data in open access repositories, although this process is still in its initial stages, given the variety of organizations involved (universities, research centres, governmental agencies, international collaborators, etc.). While the more efficiently designed repositories use standardized formats to export content, others cannot be easily integrated into the system because they do not follow interoperability protocols. The creation of repositories that can be integrated in a single network is therefore key in the collection, preservation, dissemination and reuse of research data.

Finally, although there are studies on specific international projects, such as Levitus et al. [5] on the structure of the World Ocean Database Project (produced by the US National Oceanographic Data Center, which is funded by the US National Oceanic and Atmospheric Administration and by the National Centers for Environmental Information), there is still a lack of literature on the state of oceanographic data repositories across different countries, which the present paper intends to remedy.

2. Materials and Methods

This paper reviews the current situation of oceanographic data repositories across different countries and evaluates them according to a series of indicators, including geographic location, data type and format, query system, charges and usage rights. We have adapted a complete set of indicators for general repositories presented by Serrano et al. [6].

There were three phases in our methodology: first, we located and selected a number of repositories; second, we established a series of indicators; and third, we used the indicators to evaluate each repository. We established an external connection to the web server and conducted several searches to complete the set of indicators established. We made no use of staff questionnaires. The repositories selected for the study included data providers with aggregators. This feature will be elaborated upon in the Results section.

We located and selected the repositories using the aggregators and databases of the International Ocean Discovery Program (IODP), the Woods Hole Oceanographic Institution (WHOI), OceanDocs, Aquatic Commons (IAMSLIC) and SeaDataNet. We also searched the directories of the open access registries ODiSEA (International Registry on Research Data), ROAR (Registry of Open Access Repositories), OpenDoar (Directory of Open Access Repositories), and re3data.org (Registry of Research Data Repositories). Repositories of polar data were excluded from the search.

A total of 15 national and international repositories were selected. These are described below in Table 1.

Table 1. Oceanographic data repositories.

Repository	Description
Australian Ocean Data Centre Joint Facility (AODCJF) http://portal.aodn.org.au/aodn/	Australia's designated National Oceanographic Data Centre, created in 1964 and based in Tasmania. The AODCJF is a joint venture between six Australian government marine data agencies.
Banco Nacional de Datos Oceanográficos (BNDO) http://www.mar.mil.br/dhn/chm/oceanografia/bndo.html	Brazil's designated National Oceanographic Data Centre, created in 1994 and based in Rio de Janeiro. The BNDO is supervised by the Brazilian government's Centro de Hidrografia da Marinha (CHM)
British Oceanographic Data Centre (BODC) http://www.bodc.ac.uk/	The United Kingdom's designated National Oceanographic Data Centre, created in 1969 and based in Liverpool. The BODC is funded by the National Environment Research Council (NERC).
Centro Argentino de Datos Oceanográficos, Servicio de Hidrografia Naval http://www.hidro.gov.ar	Argentina's designated National Oceanographic Data Centre, created in 1974 and coordinated by the naval hydrographic service of the Argentine Ministry of Defence, Buenos Aires.
Centro Nacional de Datos Oceanográficos do México (CeNDO) http://cendo.ens.uabc.mx/	Mexico's designated National Oceanographic Data Centre, created in 2011 and based at the Autonomous University of Baja California, Campus Ensenada (UABC).
European Marine Observation and Data Network (EMODnet) http://www.emodnet.eu/	An Ostend-based consortium of over 100 organizations making marine data available to the public, created in 2009 by the European Commission's Directorate- General for Maritime Affairs and Fisheries (DG MARE).
Geological and Geophysical Data (Geo-Seas) http://www.geo-seas.eu/	A Nottingham-based initiative of the European Commission's Research Infrastructures programme (20092013) which continues to offer data and services to the public. Steered by the project coordinator NERCBGS (the Natural Environment Research Council and the British Geological Survey) and the technical coordinator MARIS BV (Marine Information Service, the Hague, Netherlands), Geo-Seas brings together 26 data centres located in 17 European maritime nations.
Integrated Marine Observing System (IMOS) http://imos.org.au/home.html	An initiative led by the University of Tasmania in partnership with ten other academic institutions, created in 2007 and supported under the Australian government's National Collaborative Research Infrastructure Strategy (NCRIS).
JERICO http://www.jerico-fp7.eu/datatool/	A consortium of 27 partners created in 2011 and coordinated by the Institut français de recherche pour l'exploitation de la mer (Ifremer).
MyOcean www.myocean.eu/	A consortium of 59 partners in 28 countries created in 2009 and directed by Mercator Océan (Ifremer).
National Oceanic and Atmospheric Administration (NOAA) http://www.nodc.noaa.gov	The United States' designated National Oceanographic Data Centre, created in 1961 and based in Silver Spring, Maryland.
Ocean Data Portal (ODP) http://www.oceandataportal.org/	A portal created in 2013 by the International Oceanographic Data and Information Exchange (IODE) programme of UNESCO's Intergovernmental Oceanographic Commission (IOC-UNESCO) through the network of National Oceanographic Data Centres.
Rolling Deck to Repository (R2R) http://www.rvdata.us/	A programme created in 2008 and coordinated by the University-National Oceanographic Laboratory System (UNOLS) at the University of Rhode Island Graduate School of Oceanography.
SeaDataNet http://www.seadatanet.org/Data-Access	A consortium of 49 partners in 35 countries, created in 2004 and coordinated by the Institut français de recherche pour l'exploitation de la mer (Ifremer) in Brest, Brittany; technical coordination by MARIS BV (Marine Information Service, the Hague, Netherlands).

Our final selection also excluded 'data rescue' projects such as GODAR, whose objective is to locate and digitize printed or e-data that are at risk of being lost due to media decay [7].

We then established a series of indicators to evaluate the repositories, described below in Table 2. Indicators for the analysis of data repositories.

Table 2. Indicators for the analysis of data repositories.

Indicator	Description	Values
Location	Country where the repository is located.	Name of the country
Institution	Indication of the organisms responsible for collecting data.	Name of the institution
Type of service	How the repository operates and is structured, whether by means of a data provider (who directly incorporates the data into a repository) or data aggregator (who collects data from other repositories).	Provider, aggregator
Type of data	Indication of the characteristics of the data collected.	Physical, chemical, geological, biological.
Import and export format	File formats in the data repositories analyzed.	xml, netCDF, odv, hdf, bufr, other formats.
Type of distribution	System for disseminating data.	Optical disc, Web, FTP, etc.
Metadata schema	Standards or categories used in metadata for describing content.	DC, etc.
Query system	Type of query system used and search features offered for accessing network packets.	Free text search, search by metadata, no query system.
Interoperability	Protocols observed to facilitate interoperability.	ISO, HTTP, FTP OGC, WMS, WFS, CS-W, SWE, OpenSearch, OpenID, Shibboleth, etc.
Identifier	Inclusion of a persistent digital object identifier.	DOI, ARK, XRI, Handle, LSID, OID, PURL, URI/ URN/URL, UUID
Data access policy	Description of the access policy implemented.	Open access, semi-open access, open access with prior user registration, private.
Intellectual property rights	Specification about intellectual property rights.	Type of license used
Charges	Whether users pay to access data.	Free of charge, subscription, pay for data, etc.
Statistics	Inclusion of usage statistics.	Inclusion of usage statistics.

Finally, we evaluated the databases according to the indicators we had established. All our information was obtained through external queries, even though certain indicators are difficult to evaluate using this method. The data we collected, therefore, do not come from surveys conducted with the repository curators.

3. Results and Discussion

This section examines the repositories selected for the study and evaluates them, in general terms, according to the indicators we established.

3.1. Location

The vast majority of the repositories examined in this study are located in Europe and North America, whose infrastructures are well equipped to manage the exchange of data between different platforms and can even be used to collect data from research centres in Latin America and Oceania, represented here by two repositories, respectively. A total of 60% of the repositories are state organizations. These include the French Institut français de recherche pour l'exploitation de la mer (Ifremer), the Netherlands' MARIS BV, the UK's British Oceanographic Data Centre (funded by the National Environment Research Council) and the US National Oceanic and Atmospheric

Administration. The remaining 40% are located within international organizations, such as the Oceanographic Data and Information Exchange (IODE) programme of UNESCO's Intergovernmental Oceanographic Commission (IOC- UNESCO) or the European Commission's Directorate-General for Maritime Affairs and Fisheries (DG MARE).

Notable roles are played by Ifremer, which participates in two consortia and in one National Oceanographic Data Centre, and by MARIS, which directly participates in the development of two consortia. Australia's presence is important in the infrastructure of two oceanographic data management agencies.

3.2. Type of Service

Four repositories (29% of the total) act as data providers: Rolling Deck to Repository; the Centro Argentino de Datos Oceanográficos (CEADO); the National Oceanic and Atmospheric Administration (NOAA); and the Banco Nacional de Datos Oceanográficos (BNDO). The rest are data aggregators who collect data from different repositories in order to offer more comprehensive query results and who also provide common data indexing formats and applications to facilitate data exchange.

3.3. Type of Data

By and large, the repositories contain data in all the branches of oceanographic science: physical oceanography (93.3% of the total number of repositories in the study); chemical oceanography (again, 93.3%); marine geology (100%); and marine biology (86.6%). Biological data are not used in the pan-European Geo-Seas platform, which is dedicated exclusively to geological research, or in BNDO, which is managed by the Brazilian military and specializes in hydrography.

3.4. Import and Export Format

There is a broad variety of formats amongst the repositories. Of a total of forty different formats, the most frequently used are NetCDF (in 73.3% of the repositories studied), ASCII (33%), XML (20%), ODV (20%) and CSV (20%). Approximately 300 formats are used to manage marine data formats, ingest data by the repositories or delivery formats. Note, however, that the main advantage of having so many standards is that users are offered a wider range of formats to choose from. Furthermore, users are often not offered much choice for a given data set and must spend considerable time converting between formats.

3.5. Type of Distribution

Because the volume of data is so large and the data are by nature so decentralized, the repositories in this study use a variety of methods to manage distribution. A total of 93% allow users to download the data from a webpage or portal, 40% use FTP, 20% use email, 20% use CD-ROM and 20% use DVD.

3.6. Metadata Schema

Repositories need to ensure that their data are adequately registered, managed and preserved. To do this, the datasets need be accompanied by an adequate range of schema or categories of information about how the data have been obtained (temporal or spatial information, methods and instruments used for collecting data, etc.); the scientific area they are useful for; who created them; who owns them; and how they may be reused. Schema should also include categories on quality control, specifying whether this is ensured by peer review (as in the validation of scientific literature). Together with technological interoperability, the use of adequate and standardized metadata schema is essential for the access to and reuse of scientific data [8].

Not all of the repositories in our study offer metadata schema. Where present, they generally cover such categories as scientific area (physics, chemistry, geophysics, geology and biology), format (numeric and text) and geographic area. Our analysis also showed that the repositories adopt

the Intergovernmental Oceanographic Commission's Ocean Data Standards recommendation to adopt ISO 8601: 2004 to represent date and time (YYYYMMDD and hhmmss, respectively).

Most of the repositories use a netCDF interface to describe the collections of datasets. The most common netCDF format is a single file with two parts: a header, which contains all the information about dimensions, attributes and variables, except for the variable data; and a data part, comprising fixed-size data (containing the data for variables that do not have an unlimited dimension) and variable-size data (containing the data for variables that have an unlimited dimension). Files created with the netCDF-4 format have access to an enhanced data model, which includes named groups, thus allowing data types to be defined by the user and stored hierarchically in groups (like the folders in a system of files).

The US designated National Oceanographic Data Center (NODC-NOAA) uses the metadata standard Directory Interchange Format (DIF), which comprises eight mandatory elements providing specific information on the data and up to 28 other elements that broaden and clarify the information. (Note that this standard is also used in research datasets on the Antarctic).

3.7. Query System

Fourteen of the 15 repositories only allow queries using the metadata that they have defined. The exception is Ocean Data Portal, where the user can directly search the database using identifiers of a dataset. Similarly, most repositories do not allow the user to have direct access to any file, while some use the CSW standard (Catalog Service for the Web) as a resource to specify a design pattern for the definition of interfaces for publishing and searching for collections of descriptive information (metadata) on geospatial data, services and related information objects. All of the repositories offer simple and advanced searches and search support using thesauruses, specific directories and areas of oceanographic research (bathymetry, geology, etc.), except for Rolling Deck to Repository, which has no search interface and only offers information about its cruise data. The query system used by the Centro Nacional de Datos Oceanográficos de México is unusual in that it offers results of general reports on the Datasets by subject, accompanied by descriptive reports, instead of making the datasets themselves available. The Centro Argentino de Datos Oceanográficos and Brazil's Banco Nacional de Datos Oceanográficos do not offer an online search and only answer enquiries made by email or telephone. The Australian Ocean Data Centre Joint Facility uses a geographic atlas to offer users a map interface of the region in their query.

3.8. Interoperability

The collections of digital maps and the datasets with complementary tables, figures and information which systematically describe the coast in order to manage and plan coastal regions, are often used with cartographic instruments and support materials, all of which are delivered via the internet (in FTP and HTTP protocols) or by request. There is an index based on ISO 19115 for data on individual samples, cores and geophysical measurements, and there is a single interface to access these datasets online.

3.9. Identifier

France's designated National Oceanographic Data Centre (SISMER), Rolling Deck to Repository (R2R) and British Oceanographic Data Centre (BODC) can be accessed via the Ifremer portal, which stipulates that in order to give accountability to the data producer and facilitate the citation, users can associate a DOI (digital object identifier) with any dataset. No other repository in the study offered this possibility.

3.10. Data Access Policy

Users have open access in 40% of the repositories in this study, and a limitation derived from the need for prior registration occurs in only 27%. In 20% of the repositories, certain data are subject to specific access and usage conditions (semi-open access) and in 13% the information is unavailable.

3.11. Intellectual Property Rights

In 33% of the repositories, the producers (universities, national oceanographic data centres, industrial and commercial state organizations) hold the intellectual property rights to create, share and reuse the data. Again in 33%, the infrastructures themselves have specific copyrights on the material. In 27%, the identity of the right holders is not specified and in just one case (SISMER), access and usage is regulated under a Creative Commons Attribution-Share Alike license.

3.12. Charges

Users are offered free-of-charge access in 53% of the repositories and are required to pay a fee in 27%, although this fee varies. The repositories charge for manual intervention, users pay third-party developers and there are charges for CD-ROM/DVD products, data for the client and file data, and marginal costs for scientific use. Twenty per cent of the platforms do not offer information.

3.13. Statística

The French SISMER is the only repository that offers information on statistics on data access with annual reports. The consortium MyOcean offers statistics by search area at any moment and SeaDataNet offers the DIVA software tool (Data-Interpolating Variational Analysis), which allows the user to spatially analyze observations on a regular grid in an optimal way. These gridded fields have various applications: they can be used to verify the consistency of readings (i.e., detect atypical values) and support the initialization, calibration and validation of data models (for projects such as MyOcean), analyze the changes and tendencies in seasonal, annual and interannual time scales and also analyze estimates (e.g., heat content and total biomass). These two Ifremer-coordinated repositories offer precise and detailed statistics using free-of-charge key-based authentication. The rest do not offer information about the statistical control of data.

4. Conclusions

Our analysis indicates that, by and large, oceanographic data repositories are developing their systems for processing, disseminating and reusing data. It also reflects the scientific community's growing interest in such repositories, which play an essential role in the storage and reuse of oceanographic data. As observed by Belter [4], "Data repositories creating such products could then become central hubs for disciplinary, and potentially interdisciplinary, research, leveraging the limited research funding available in each discipline to ensure that individual pieces of research performed in that discipline eventually benefit the entire disciplinary community".

Most of the repositories analyzed in this study are located in Europe and North America. Many are international organizations, although the continents of Asia and Oceania are under-represented. Many are curated by aggregators, who regularly monitor the activities of data providers to secure new content. In this way, the collection and combination of data from multiple sources can offer users a more holistic perspective of the informational environment.

The high potential of the data in these repositories can be attributed to the balance and variety in data type, meaning that although they are not always well managed or accompanied by quality assessment tools, the data do cater to the needs of various user communities, including physical oceanographers and marine chemists, geologists and biologists.

Bechini and Vetrano [9] offer a global vision of the various steps (phases) related to oceanographic data, from collecting data to publishing them or including them in repositories. They describe the difficulties of managing heterogeneous oceanographic data and the ways this can be addressed.

The query systems used by these repositories are undermined by the rigidity that characterizes metadata schema, which is widely used by all of the platforms. Only three of the repositories allow free text searches. However, in two of these three, SeaDataNet and the British Oceanographic Data Centre, the protocols controlling and assessing the quality of the metadata guide queries towards error-free syntax, in accordance with the metadata schema being used on the website. Only Ocean Data Portal, which observes the data standards of the International Oceanographic Data and Information Exchange, allows users to create a natural language query phrase or sentence that describes the subject they are looking for. The objective is to enhance the identity of the data, make them more available and easier to reach and, above all, ensure that data are more thoroughly documented.

Finally, if oceanographic data repository curators wish to improve their management systems they will need the support of the scientific community's researchers and they will also need to improve the degree of coordination between the institutions that collect and disseminate data. The repositories will also need to be more widely integrated in the work of researchers, who often fail to disseminate the data from their findings properly. (It is still common to find researchers who do not send data to a repository even when they are aware that an appropriate repository exists). However, while this may reflect the community's low level of interest in publishing data in open access media or its reticence to publish, the various synergies between scientific journals, repositories and journals can also offer valuable opportunities for discovering and reusing research results. A good example is the world atlas of marine ecosystem data MAREDAT [10,11]. In this respect, the French infrastructure for marine spatial data is an example of good practice [12,13].

We should remember that the greatest challenge facing the scientific community is to create effective data management systems for oceanographic data (collecting and processing data, making it available and reusable, etc.) and that the management of the data repositories will be essential for the success of this process.

Finally, the results obtained with the analysis of the repositories may be of interest to other disciplines, such as biology, geology, geography or other areas that have lines of research addressing marine studies directly or indirectly.

Author Contributions: Ernest Abadal conceived and designed the analysis; Fabiano da Silva performed the analysis and analyzed the data; Enrique Wulff contributed with analysis tools; Fabiano da Silva wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thanos, C. A vision for global research data infrastructures. *Data Sci. J.* **2013**, *12*, 71–90. [CrossRef]
2. PARSE. Insight. Available online: http://www.parse-insight.eu/ (accessed on 25 October 2016).
3. Research Data Alliance. Available online: https://www.rd-alliance.org/ (accessed on 25 October 2016).
4. Belter, C.W. Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE* **2014**, *9*, e92590. [CrossRef] [PubMed]
5. Levitus, S.; Antonov, J.I.; Baranova, O.K.; Boyer, T.P.; Coleman, C.L.; Garcia, H.E.; Grodsky, A.I.; Johnson, D.R.; Locarnini, R.A.; Mishonov, A.V.; et al. The World Ocean Database. *Data Sci. J.* **2013**, *12*, 229–234. [CrossRef]
6. Serrano, R.; Melero, R.; Abadal, E. Indicadores para la evaluación de repositorios institucionales de acceso abierto. *An. Document.* **2017**, *20*. Available online: http://revistas.um.es/analesdoc/article/view/analesdoc.17.2.190821 (accessed on 25 October 2016).
7. Levitus, S. The UNESCO-IOC-IODE "Global Oceanographic Data Archeology and Rescue" (GODAR) Project and World Ocean Database projects. *Data Sci. J.* **2012**, *11*, 46–71. [CrossRef]

8.　Schaap, D.M.A.; Glaves, H. ODIP—Ocean data interoperability platform—Developing interoperability pilot project 1. *Geophys. Res. Abstr.* **2014**, *16*, 17–22.

9.　Bechini, A.; Vetrano, A. Management and storage of in situ oceanographic data: An ECM-based approach. *Inf. Syst.* **2013**, *38*, 351–368. [CrossRef]

10.　Buitenhuis, E.T.; Vogt, M.; Moriarty, R.; Bednaršek, N.; Doney, S.C.; Leblanc, K.; Le Quéré, C.; Luo, Y.-W.; O'Brien, C.; O'Brien, T.; et al. MAREDAT: Towards a world atlas of MARine Ecosystem DATa. *Earth Syst. Sci. Data* **2013**, *5*, 227–239. [CrossRef]

11.　Quadt, F.; Düsterhus, A.; Höck, H.; Lautenschlager, M.; Hense, A.; Hense, A.; Dames, M. Atarrabi: A workflow system for the publication of environmental data. *Data Sci. J.* **2012**, *11*, 89–109. [CrossRef]

12.　Satra-Le Bris, C.; Quimbert, E.; Treguer, M.; Louarit, A. Sextant: French spatial data infrastructure for marine environments. In Proceedings of the International Conference on Marine Data and Information Systems (IMDIS 2013), Lucca, Italia, 23–25 September 2013.

13.　Merceur, F. Référencement et DOIfication des jeux de données géographiques. In *Journée "Administrateurs Sextant"*; IFREMER: Plouzané, France, 2014.