Noname manuscript No.
(will be inserted by the editor)

Facilitating Open Exchange of Data and Information

James Gallagher · John Orcutt · Pauline Simpson ·
Dawn Wright · Jay Pearlman · Lisa Raymond

Received: date / Accepted: date

Abstract

By broad consensus, Open Data presents great value. However, beyond that simple statement, there are a number of complex, and sometimes contentious, issues that the science community must address. In this review, we examine the current state of the core issues of Open Data with the unique perspective and use cases of the ocean science community: interoperability; discovery and access; quality and fitness for purpose; and sustainability. The topics of Governance and Data Publication are also examined in detail. Each of the areas covered are, by themselves, complex and the approaches to the issues under consideration are often at odds with each other. Any comprehensive policy on Open Data will require compromises that are best resolved by broad community input. In the final section of the review, we provide recommendations that serve as a starting point for these discussions.

Keywords: Open Data · Interoperability · Governance · Data Publication

Corresponding author: J. Gallagher
OPeNDAP, Inc. Narragansett, RI. E-mail: jgallagher@opendap.org, phone: 401.575.3296

J. Orcutt
Scripps Institution of Oceanography/University of California, San Diego. E-mail: jorcutt@ucsd.edu

P. Simpson
Central Caribbean Marine Institute, Cayman Islands. E-mail: psimpson@reefresearch.org

D.J. Wright
Environmental Systems Research Institute, Redlands, CA. E-mail: dwright@esri.com

J. S Pearlman
University of Colorado, Boulder, Colorado. E-mail: jay.pearlman@ieee.org

L. Raymond
MBLWHOI Library, Woods Hole Oceanographic Institution. E-mail: lraymond@whoi.edu

1 Introduction

There is great interest in the idea of Open Data and the exploitation of such data for a wide range of purposes. Open data has particular impact in sciences that are integrative and are collaborations across disciplines and sub-disciplines (Carpenter 2009). While this paper focuses on the ocean sciences, the issues, in general, are common across disciplines including biology (Thessen 2011) ecology (Reichman 2011) and others where the diversity of data and acquisition/documentation processes vary widely. Even the definition of the term "data" varies widely across dialogues of open data (Thessen 2011). The term "data" used in the context of this paper is broad; data are factual information used as a basis for reasoning, discussion, or calculation and the term is not limited to Internet modalities. The definition of "data" covers visualizations, analyses, model outputs and the underlying digital or other information that may be used for analysis and other functions. Open Data, as defined in the Open Data Handbook (Open Knowledge Foundation 2012) are "data that can be freely used, reused and redistributed by anyone—subject only, at most, to the requirement to attribute and share alike." The handbook expands on the definition:

– Availability and Access: The data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the Internet. The data must also be available in a convenient and modifiable form.
– Reuse and Redistribution: The data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets.
– Universal Participation: Everyone must be able to use, reuse and redistribute—there should be no discrimination against fields of endeavor or against persons or groups. For example, 'noncommercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.
– With increasing access to data through high speed Internet, Open Data for research and government has become an important area of discussion and debate.

Looking at the science community, Costello (2009) and Borgman (2012) list many benefits of sharing data in the sciences for:

- Individual scientists, both as data creator and as researcher and author: Additional publications; Greater citation rate (Piwowar 2007); and Wider recognition among peers.
- Editors and reviewers: Independent verification and qualification of research findings
- Publishers: Citation of data publications is likely to increase citations of related research papers.
- Data centers: Increased value and role in science.
- Scientific community: Reuse data and integrate data with other data to create new resources; Reproduce or verify research; and Enable others to ask new questions of extant data.
- Funding agencies: Better financial return from research investment as data can be used again.
- Governments: Data accessible to government science advisors.
- Society: Better science; Make results of publically funded research available to the public; and Advance and accelerate the state of research and innovation.

The identification of Open Data is an issue that is recognized globally. The governments of the European Union, the United States, Japan and Australia have all made open data a critical part of their policy. In a 2011 press release, EU Commissioner for the Digital Agenda, Neelie Kroes, noted that the EU public sector is "sitting on a goldmine of unrealized economic potential expected to deliver a €40 billion boost to the EU's economy each year. … To achieve this potential, data must be accessible and open." (European Commission 2011) Similar statements by representatives of the US, Japanese and Australian governments mirror this sentiment (NSF 2014; Guess 2013; Obama 2013; Australian Government 2009).

While the directions are clear, the implementation of an Open Data environment has its own challenges and currently the science community has large variations in the openness of its data. The perceptions of an emerging modern deluge of data in both government and science (Borgman 2012) stands in stark contrast to the lack of progress of sharing both within and across government agencies and across science disciplines. The challenge is not only a technical one. The evolution will require changes in our cultural system of rewarding scientist and engineers for their innovation. Neilsen (2011) pointed out that the benefits of Open Data are not only an increase in data availability, but also a cultural change resulting in an interesting new approach to the conduct of science:

“The reinvention of discovery is one of the great changes of our time. To historians looking back a hundred years from now, there will be two eras of science: pre-network science, and networked science. We are living in a time of transition to the second era of science. But it’s going to be a bumpy transition, and there is a possibility it will fail or fall short of its potential.”

Throughout this paper, we highlight the unique perspective of ocean science, which is interdisciplinary in its nature and requires understanding of not just one basic science but a combined knowledge of biology, chemistry, physics, geology, geophysics, and engineering, in order to fully understand patterns and processes in the oceans. This also requires a multidisciplinary understanding of the collection, formatting, and open provisioning of the data in those sub-disciplines.

2 Challenges for an Open Data Environment

The philosophy of free and open exchange of ideas and information has long been a tradition of science, but the extension of these principles to raw data or comprehensive calibrated data, along with the term “Open Data” itself, is relatively new (Vision 2010). It follows many other “Open” concepts such as open source software and has been adopted by governments to suggest transparency in their operations. Even with the tradition of openness in science, moving forward to a uniform approach for Open Data is not straightforward. There are a variety of approaches to collecting environmental data ranging from single-investigator field experiments, which may last for only a short time; descriptive programs that are conducted for civil purposes (e.g., beach quality or oil spills) and observatory systems with a goal of collecting data over a long period of time. These different approaches to data collection have widely divergent resources available for data management (e.g., integrity and/or quality control, provenance, metadata definition, timing and others) so that a single solution for opening data to external access is unlikely. Similarly, the translation from data to information and then knowledge through models involves techniques that range from single-focus analyses to community models. Moving from community or discipline-specific models to global simulation and prediction is yet another step in complexity and further motivates the need for Open Data and access.

In a survey and analysis of open-data repositories (Braunschweig et al. 2012), two major problems common to almost all platforms were dead links and a plethora of different file formats. Web services address the problems of file transfers by hiding the actual file format. They also provide subsetting and aggregation to reduce the quantity of data transferred. Even with this reduction in transfer, they still do not completely address network bandwidth issues and they still can be technically challenging for some users. The technical challenges can be reduced through techniques improving commonality of descriptions and through even broader interoperability approaches. There are many levels of interoperability from basic machine interactions to human exchanges to human rewards and motivations.

On the machine side, two extremes have been identified and there are a variety of approaches that mix varying degrees of each of them. The first is to provide an intermediary information system layer that translates between different domain information infrastructures allowing the domain system to maintain its independence while enabling full interoperability (Nativi et al 2012). The second approach is to mandate certain standards that must be followed by each domain system so that the different systems will be interoperable (Busse et al 1999). The former is a brokering approach and the latter, a federated approach. Both of these must ultimately address the issues of semantics, metadata, workflows, and so on. The brokering approach reduces the workload on discipline repositories by centralizing the interoperability developments into the middleware layer. This encourages greater participation on the part of the discipline information infrastructures by reducing local efforts.

For the human side, the cultural issues represent significant challenges. These include academic recognition and promotion for collection and publication of data. The rationale for protecting data from external view stems largely from the academic rewards systems in which scientists are judged by their analyses published in papers as well as the number and quality of subsequent references to the work. The ultimate academic goal is writing a scientific paper, which can lead to increases in salary and grants. Were the data open, competitors could gain an “unfair advantage” because they do not have to do the original work in data collection or data processing. In a study on the willingness to share, D.S. Sayogo showed that reward was found to have a significant indirect impact on data sharing, which leads to the issue of considering how to define rewards to encourage sharing behavior in collaboration (Sayogo and Pardo 2012). Only recently has there been provision through the use of Digital Object Identifiers to enable effective referencing of data sets.

Even when scientists do want to make their supplementary research material available, such as software and mathematical proofs, they may need assistance in doing so. In a study of linking data to publications, a project was done to help researchers link their datasets to their publications, thus creating “enhanced publications.” (SURF 2013). This issue could be effectively addressed through documentation of accepted practices (“best” practices) that can be referenced by for approaches to data release, formats, languages or semantics, quality assessments and communication protocols (Whitlock 2011; Costello and Wieczorek 2013).

3 Core Issues for Open Data

For implementation, there are core issues for Open Data that flow from the desire to use Open Data for new and sustainable applications. These core issues for Open Data are:

1. Ability for data to be discovered, accessed, and used across domains with different cultural backgrounds;
2. Transparency and information supporting use such as quality and fitness for purpose (i.e., data integrity); and
3. Sustainability for future access.

There is overlap in the above issues and the boundaries are indistinct. Thus the discussion below, although formatted in the context of the above three issues, must be thought of in the context of the overall challenge of using and benefiting from access to Open Data. From this perspective, the core issues are addressing various facets of long-term interoperability. Open Data should support interoperability between domains and between communities for it to have the broadest utility.

3.1 Discovering and Accessing Data

Access to open data using the Internet has multiple facets including machine-to-machine file transfers and query-based data retrievals from specialized data servers. Of course, file transfers are technically ‘query-based retrievals’ since the files must be requested (the query) and sent from the source machine using some sort of software program (a server). However, there are differences between the two cases. Static data files hold a predetermined package of data whose make up was determined prior to any given user’s request for those data while specialized data servers typically implement query and transfer protocols that provide a way to transform data before it is sent to the requestor such as selecting certain geographic boundaries for the retrieval and transferring only the appropriate data.

Data access using simple file transfer over the Internet is often accomplished using File Transfer Protocol (FTP) or Hypertext Transfer Protocol (HTTP), although this short list is not exclusive.[1] FTP, HTTP and other protocols provide ways to navigate remote file systems and transfer files. FTP provides features for automating the process to some degree while HTTP, which is the transport protocol

[1] e.g., file synchronization protocols like rsync could be used.

used by the Web, has the advantage that it's widely supported (every host on the Web has a HTTP server running and almost every Internet user has, and is familiar with, a web browser). Both protocols support anonymous and authenticated access as well as logging all accesses. More recently FTP has been generalized to GridFTP to support more reliable and high performance file transfers. The new protocol makes better use of available bandwidth (e.g. 10Gbps and higher) by using multiple, simultaneous TCP streams (Alcock et al, 2005). The Ocean Observatories Initiative adopted a protocol based on the Advanced Messaging Queuing Protocol (AMQP) to provide secure point-to-point connections capable at operating at very high bandwidths (Marshall 21012).

While the strengths of file-based data access are significant, there are also drawbacks. Because the content of the files (i.e., the unit of transfer) is predetermined, it will not be a perfect fit for most users. Instead, users will likely need to get software to read the files, extract and transfer information from the files to some visualization or analysis tool and (often) subset those files. Beyond this, many datasets are actually stored as a set of files, and remote users must understand how those files Are combined to form the whole dataset. This knowledge is required to enable the user to correctly request a specific set of files and then read from each, combining their contents to form a coherent whole.

To address the shortcomings of file-based access, a number of other protocols have been developed that provide richer query interfaces, return data in different formats and provide remote processing capabilities. These interfaces typically are combined with, or contain as an integral component, a catalog protocol that provides a way for remote users to discover both dataset contents and the parameters that may be used to query and subset/transform those contents in a request. Typical examples are the WMS/WCS protocols developed by the OGC (Whiteside and Evans 2006), and DAP developed by OPeNDAP (Gallagher et al 2007). Using such an interface, remote users can request subsets of data custom-tailored to their specific needs, regardless of how those data are stored on the server and, for most of these protocols, in a format most suitable to their software. This provides distinct benefits over file access protocols such as FTP because users do not have to decode files in order to get just those data they need, and remote sites can retain their idiosyncratic storage formats. These benefits translate into less work for both data users and providers and a savings in network bandwidth.

Falling between simple file transfer systems and web services that hide data formats completely are systems like DataONE (Reichman 2011) which provides users with a custom web services interface to upload and manage datasets in a distributed system that handles replication and cataloging functions. DataONE combines these web services, which include access to a searchable catalog, with simple file access over HTTP when actually downloading data. Similarly OBIS (Grassle 2000) and WoRMS (Costello and Bouchet 2013) provide access to earth science data stored in one or more relational database systems but also include interfaces based on web services. The WoRMS system integrates "over 100 global, 12 regional and 4 thematic species databases with a common taxonomy" and combines a relational database accessed using a web service interface with WMS for access to maps of species distribution. Like WoRMS, OBIS is a database system (serving marine animal biogeographic data) with a web service interface that conforms to the open standards published by the OGC (Grassle 2000; Best 2007). Each of DataONE, WoRMS and OBIS combine web services with file transfer or database access to provide flexible online systems.

3.3 Data Quality and Fitness for Purpose

Ultimately, fitness for purpose is a key attribute that must be understood to use data. This includes factors such as data quality. Challenges are inherent in the increasing diversity of data resulting from the introduction of new technologies in observation and communication. Citizen science introduces data that can have large differences in quality due to the difference in expertise of observers. Even automated instruments can introduce unknown variations due to external noise, poor timing, biofouling or uncertainty due to a sampling process (e.g., signal conditioning). When data from one oceanography discipline (such as the ocean surface temperature) is combined with data from another (fish abundance), the uncertainties in the combined data may not be as easily quantifiable as that of the individual contributing data sets. There are issues that a nonexpert user may not even consider. For example, timing can be an issue when it's important to combine data series (e.g. pressure and conductivity) and the accuracy of the clocks is unknown. The time in one time series can differ by many seconds from the other; coherence studies between different observational platforms, while potentially of importance for understanding the underlying transfer function, may be impossible.

These uncertainty issues become more important when data are freely distributed to users with diverse interests and skills. A way to address this is through the adoption of Open Data quality indicators. The primary level of quality indicators might be a flag indicating good, bad, missing data, and data that are questionable because they fail some non-critical test. Secondary quality designations can be more specific and vary by data type. Excessive gradient, excessive spikes, unexpected ratio of observations, and many other data quality tests can be applied at this secondary level and the flags stored with the data (for example, Folkman, et al 2003). Various international projects are looking at quality indicators. CEOS QA4EO (GEO/CEOS 2008; Lecomte and Stensaas 2009) is a quality assurance protocol from the Global Earth Observation System of Systems (GEOSS) (Pearlman and Shibasaki 2008). GeoViQua (2007) has focused on adding rigorous quality specifications to the GEOSS spatial data in order to improve reliability for scientific studies and policy decision-making. For real time data, quality assurance is more challenging because the quality process must be automated and be robust. Quality Assurance of Real Time Ocean Data (QARTOD), a component of the U.S. Integrated Ocean Observing System (IOOS) (NOAA IOOS 2014) addresses these challenging data quality issues.

Provenance and traceability support knowledge of uncertainty in the data and are another important element in gauging the fitness for purpose of data and information. "While 'fitness for purpose' is the principle universally accepted among scientists as the correct approach to obtaining data of appropriate quality, many scientists or end-users of data are not in a position to specify exactly what quality of data are required for a specific analysis" (Whitfield 2012). This is a particular problem in long-term studies where the data are produced by a multitude of sensors that may not be cross-calibrated. Generally, agencies collecting environmental observations provide data "as is" with no warranty as to its fitness for any particular purpose even when they assess observation errors. Since fitness for purpose is in the eye of the beholder, there is, in fact, no quantitative metric that can be applied uniformly.

3.4 Sustainability

Sustainability of the Open Data paradigm is a major issue and one of the still unanswered questions in the move toward open data. Sustainability, or the ability of the Open Data approaches to be maintained over time, involves a combination of resources, human factors and policy. The evolution supporting sustainability may take more than a generation.

Data and its provenance must be preserved over the long-term along with the associated software that apply to the data and its analyses. Ownership of the data and the innovations fostered are embedded in intellectual property rights (IPR) that govern who benefits. The IPR laws regarding ownership of the outcomes of scientific research in the US changed to allow universities to retain the IPR, and this has become a significant business opportunity for educational institutions. Publishing houses, both profit and non-profit, including science and technical organizations (IEEE, AGU, AAAS, etc.) often retain the copyright for all articles they publish, selling subscriptions and access to their resources through subscriptions to university libraries and others. In return for this resource base, publishing houses make an important contribution to the quality of the scientific literature by running the peer review process and management of repositories. The peer review system, although postulated to allow replication of scientific discoveries, did not require that data used in analysis for such publications be released. In the academic culture, data publication was not considered strongly for decisions on tenure track and promotions.

In the move toward open data, many of these issues and the financial impacts of changes mean a restructuring of the business models and individual incentives within the current research environment. It also raises complex questions about the ownership and rights for non-digital data such as biological specimens or rock samples. The National Science Board held a study on this subject, raising these questions and many more relating to the management, business models and rights with respect to Open Data (NSB 2011). The task force on data policies recognized that a key challenge with respect to longevity and sustainability is in the uncertainty for support of the full data life cycle: “Data stewardship is critical to the longevity and sustainability of data sharing and management throughout the data lifecycle, but it is unclear where the responsibilities for this effort lie.” In their recommendations, they recognized that “Stakeholder roles, responsibilities, and resources must be clearly identified and proactively established to support sharing, management, preservation, and long-term digital research data accessibility” and recommended the formation of a panel of stakeholders “to explore and develop a range of viable long-term business models and issues related to maintaining digital data and provide a key set of recommendations for action.” Furthermore, Costello et al. (2014) conclude that, at least in the subject area of biodiversity, larger databases have a greater likelihood of being sustained and preserved than smaller ones. They also note that if “databases are owned and curated by a collaborative partnership including a science organization … with a suitable mandate” then sustainability of the database is more likely. T. Vision (2010) recommended a similar model noting that large infrastructure/facilities will be in a better position to address long-term sustainability.

While the core technical capabilities exist for managing Open Data, there are financial and policy issues that have yet to be addressed by the National Science Foundation. Agencies in the US and governments outside the US are creating or modifying their own policies with potentially important variations in the implementation details. Thus, leadership in implementation approaches and ultimately consistency across government organizations is a critical step in providing sustainability of Open Data.

4 Uses Not Intended—A Benefit from Interoperability

The innovation and new information that stems from an Open Data paradigm comes, in part, from data being used in a wider range of applications than originally envisioned—uses that were not the intent of the original scientific observation or analyses. The rising tide of globally available digital data will create many such opportunities for science and for society, but the data need to be harnessed by a new breed of data infrastructures that are based not only on the interoperability of systems but also the interoperability of multiple disciplines in the physical and social sciences, engineering and the humanities. As mentioned earlier, interoperability is a foundation in addressing the Core issues discussed in Section 3 above. In recent years, important programs and initiatives are focusing on this challenge, including:

- In the European Union: The European Infrastructure for Spatial Information in the European Community INSPIRE (2014), and the Global Monitoring for Environment and Security (GMES) (Copernicus 2014);
- In the United States: The US National Spatial Data Infrastructure (NSDI) (FGDC 2014), Data Observation Network for Earth (DataONE 2014) and the recent EarthCube (NSF 2014) and;
- Internationally: The international initiatives Global Earth Observation System of Systems (GEOSS) (Pearlman and Shibasaki 2008).

There are several well-known disciplinary infrastructures, such as: WMO Information system (World Meteorological Organization 2014), the Global Biodiversity Information Facility (GBIF 2014), the Ocean Biogeographic Information System (OBIS) (Grassle, 2000), the Pan-European Infrastructure for Ocean & Marine Data Management (SeaDataNet 2014), the US CUAHSI Hydrologic Information System (CUAHSI 2013), the IODE infrastructure for oceanographic data and information exchange (IODE 2014), the Incorporated Research Institutions for Seismology (IRIS 2014) and a global geology information network, OneGeology (OneGeology 2014). There are others under development, including: the European Plate Observing System (EPOS) (Cocco 2012) and the GEO Biodiversity Observation Network (GEO BON) (Earth Observations 2013).

According to a study of the European Commission (EU 2006), interoperability encompasses at least three overarching and different aspects:

1. Semantics, which ensures that exchanged information is understandable and usable by any application or user involved;
2. Technology, which concerns the technical issues of linking up computer and information systems, the definition of open interfaces, data formats and protocols.

3. Organization, which deals with organizational processes, aligning information architectures with organizational goals, and helping these processes to co-operate. This category can also include important interoperability challenges, like: data policy, legal, cultural, and people harmonization.

Interoperability is not an on-off capability; there are various levels of interoperability. Different models for levels of interoperability already exist and are used successfully to determine the degree of interoperability implemented by a disciplinary infrastructure. One of them: the Levels of Conceptual Interoperability Model (LCIM) applies well to assess the Earth Sciences infrastructure levels of interoperability (Turnitsa 2005). This goes beyond the technical interoperability addressing conceptual/semantic models interoperability. The seven layers of the LCIM provide a finer granularity view of the first three levels in the model by Palfrey and Gasser (2012) previously described and are as follows:

Level 0 (No Interoperability) Stand-alone systems–no data are shared.
Level 1 (Technical Interoperability) A communication infrastructure is established, underlying networks and communication protocols are unambiguously defined.
Level 2 (Syntactic Interoperability) A common protocol to structure the data is used; the format of the information exchange is unambiguously defined.
Level 3 (Semantic Interoperability) The meaning of the data is shared through the use of a common reference model and the content of the information exchange requests are unambiguously defined.
Level 4 (Pragmatic Interoperability) The meaning of the data and the context of their use are “understood” by the participating systems, and the context in which they are exchanged is unambiguously defined.
Level 5 (Dynamic Interoperability) Systems are able to comprehend the state changes that occur in each other system’s assumptions and constraints over time; thus, the effect of the information exchange is unambiguously defined.
Level 6 (Conceptual Interoperability) The conceptual models underlying the data in each system are aligned. This requires that conceptual models be documented so that other engineers can implement them using only their specification.

Standards are essential to both machine-to-machine and data-level interoperability. A range of technologies is needed to realize even a simple interoperability framework because no one standard currently provides anywhere near the breadth of coverage needed. Instead, it is common to combine several standards to achieve a set of interoperable technologies that can work cooperatively to form a framework (Hankin et al. 2010). For example, the NSF Ocean Observatories Initiative has adopted an internal data model, which can be served to users in a variety of formats including MATLAB, CSV, ASCII, and JSON. Often these are a mix of formal and de facto standards from both formal organizations whose mission is to promote standards and grass roots community efforts. Organizations that provide a formal framework within which standards are defined and made available include IEEE, IETF, W3C, OGC, ISO and others. Standards from these organizations define the protocols used for most computer communications as well as important data format and metadata standards.

‘Community standards’ generally promote interoperability within a specific discipline at the level of interpreting content as opposed to communication protocol or format. Two examples of such standards are Darwin Core (Wieczorek et al. 2012) and Climate Forecast (Hankin et al. 2010), both of which are the product of a community effort to make data “accessible, discoverable and integrated” (Wieczorek et al. 2012). In both cases the standards meet the specific needs of the community members who work at developing them, reflecting a high level of pragmatism (Hankin et al. 2010). However, community standards do not always support the broader needs that come from research using cross-disciplinary data and information. One approach to addressing the limitations imposed by a single-community focus is to adopt a reference model such as the Open Archival Information System (OAIS). While OAIS is primarily a discipline-independent abstract model for data archives (Lavoie 2008), it necessarily addresses the concept of interoperability. However, OAIS, and similar reference models, can be used to ensure good practice but they are not a substitute for technical analysis and specification of services (Allinson 2006). Thus, convergence is a challenge in moving interoperability to higher levels and care must be taken to define the objectives for this.

Middleware, software that translates from one software protocol to another and that supports translation from one community's formats and standards to those of another community, can be used to establish interoperability. While cost of middleware can be high, the conversion of domain information systems into a single format environment is unlikely, given the additional load on the information technology teams and the likelihood that technology evolution will make any specific solution obsolete over time. In addition, interconnecting existing discipline specific systems has traditionally introduced limitations to their autonomy and scope. Because different disciplines historically have developed different approaches and technologies to collect, encode, and exchange data, bridging disciplines is a complex challenge. The brokering approach can be used to handle such differences without limiting autonomy or putting a significant investment burden on existing systems (Nativi et al 2013). The brokering approach integrates and supplements the standardization approach, building an effective system of systems out of otherwise autonomous systems. Ultimately, interoperability solutions of a global nature will be a combination of middleware (e.g., brokering) and standards (both formal and de facto).

5 Governance—Business Models and Policies

The Open Data system should be financially sustainable in order to provide continuous, long-term service. This relates not only to access to data, but an ability support those factors such as citation references and other related information that impact financial, career growth, grant selection for data/information suppliers and users. Thus, a discussion of Open Data should also deal with comparative national and international data policies, current business models for Open Data, and intellectual property rights (IPR). With respect to the ocean sciences, such a discussion raises a number of issues:

– What are the restrictions on data access and how do these impact research?
– What policy would best balance the interests of the researcher and society?
– What is the balance between Open Data and intellectual property rights?
– What are the roles of different organizational types in stimulating and funding ocean research?
– What are the data access models including IPR, business models for Open Data, data policies, and real-time assured access.
– What are the implications for security?

Borgman (2012) has contrasted some of these facets for several observational programs. One of these programs, beach quality, is a project undertaken in response to government requirements for quantifying hazards to recreational activities. On the opposite end of the scale is a study of Star Dust, generally a purely scientific endeavor, but one that can be classified as an observatory.

For the beach quality measurements, the observations entail:

Specificity of purpose:	Exploratory
Scope of Data Collection:	Describe phenomena
Approach to research:	Empirical/measurements
People involved:	Individuals
Labor to collect data:	By hand
Labor to process data:	mid-way between "By Hand" to "By Machine."

For Star Dust there are contrasting descriptions where the classifications are:

Specificity of purpose:	Observatory
Scope of Data Collection:	Model system
Approach to research:	Theoretical
People involved:	Collaborative team
Labor to collect data:	By machine
Labor to process data:	By machine

Clearly, the latter will have substantial funds available for preparing data, assigning metadata, and providing users with the data collected. For the former, the likelihood of significant funding for data access is small and records largely comprise lab notebooks (at least in southern California). Expectations for data access will be greater for the observatory and less for studies such as beach quality. Nevertheless, Open Data serve other researchers, civil purposes and the general population. Consideration must be given to mechanisms for useful data uptake even those from small programs with few resources.

For observatories with well-developed data systems for metadata definition and versioning, it's important to maintain metrics over the years for data usage to allow data sets to be "pruned" over time. While data storage costs will decrease exponentially with time, the need for persistence of some data may be questionable; for example, sensors with substantial flaws such as drift or poor timing, which simply aren't used, may be candidates for removal. As data become more open, overlap of data in repositories will be observed. The question will be asked which sets are of "higher" quality and what should be maintained. For international comparisons, national priorities will play a role in the decision process. For the ocean community, a working group of repository and cyberinfrastructure leads could support decision processes through assessment of available Open Data.

5.1 Business models

Four decades ago, the primary business model for scientific publication was subscription fees augmented (20%) with per-author charges for pages and color photographs (Björk 2012). Over the last ten years, the boundaries of this model have changed, particularly due to the rise in number of Open Access journals, which has grown at a rate of 18% (Laasko 2011). These changes are due both to the advent of inexpensive storage, the pervasiveness of high speed Internet and the impact of Open Source software on the publishing world (Björk 2012). As Björk (2012) reports, while a minority of Open Access journals require author fees, the number is still significant (approximately 25%) and may be growing; author fees were assessed by 43% of the scientific journals surveyed by Kozak and Hartley (2013). These changes are relevant to Open Data because it seems likely that similar business models will be applied to data. The models are evolving rapidly and the environment is competitive. Examples of alternatives relevant to Open Data are:

1. Amazon built an array of servers to support their online business. They now offer space on servers using "the Simple Storage Service (S3) cloud" which is available to science users and the general public. The cloud offers advantages of reliability, expandability and other attributes that have resulted in substantial use for data storage. There are a number of subscription storage models that address wide ranges of information exchange such as DropBox that serve the science community.
2. Google provides search services through its search engines and storage system. It provides visualization of scientific data through Google Earth. These services are free to users, paid for by advertising. As the market expansion for advertising revenues began to saturate, Google turned to selling focused marketing information to businesses. In some sense, Google users have given up some degree of privacy in exchange for free usage.
3. For publishers of scientific journals, as mentioned earlier, there is a transition from subscription charges to author fees often supported by government funding or the author's employer. As Kozak and Hartley (2013) point out, the number of open access journals that assess author charges varies widely by discipline from 47% (medicine), 43% (science) to 0% (the arts). Whether this will be viable in the long run (in the Open Data model) is still to be determined. Publishers are also adopting added "value" features such as the implementation of Digital Object Identifiers (DOI) so that underlying data sets for a publication, in addition to the publication itself, are identified and potentially accessible.
4. For observatories, data storage is supported by the observatory sponsor for long periods (decades). As long as the cost of the storage including its maintenance is supported (quality, provenance, etc), this is an attractive option for assuring the long-term availability and free access to data. However, the operations budget for an observatory competes with research funds and this creates a tension in the research community. NSF established a series of DataNet programs providing a decade of support. DataONE (2014), as an example, has full funding for the first five years and then

decreasing support during the next five with a transition to self-supported operation at the end of the funding decade. The Data Federation Consortium (DFC) (DataNet 2014) is a similar undertaking although starting after DataONE. The model for such a transition is not clear at the present time, particularly in an Open Data environment.

5. The Open Geospatial Consortium (OGC 2014) is an international standards organization and derives its operational costs from membership dues and from government grants for standards implementation and support. The business model comprises membership dues defined according to the type of a participating organization. The OGC business model is different than that of other standards organizations such as International Organization for Standards (ISO) and the IEEE, which charge users for standards documentation. The OGC model has been effective in rapidly responding to community needs as interoperability standards have expanded from data interoperability to sensor and model webs.

Which of these will survive the test of the marketplace and which will ultimately support Open Data sustainability is difficult to predict. The preferred outcomes of the successful business models (as there is not likely to be only one), however, can be described:

- Ensures sustainability;
- Preserves the peer review attributes of science and of publications;
- Assures scientists of recognition for their scientific research;
- Maintains data attributes such as provenance, metadata, quality attributes, etc.;
- Allows easy discovery and access to data and information, particularly supporting cross discipline research;
- Supports IPR and licensing protocols;
- Consistent with national and international policies;
- Motivates participation and contributions;
- Has minimal impact on existing disciplinary systems;
- Works across physical, social and economic sciences; and
- Accessible and usable by the public.

There likely will be a mix of systems supporting the above attributes. The uptake of the business community of these attributes will be essential, but is not guaranteed. Part of the challenge is that some of the above attributes are policy related and policies vary according to nations and in time. In particular, scientific research is predominantly government supported including publishing and data management. As the Open Data policy expands, the government funding will need to account for the different conditions and attributes of the policy. In addition to monetary resources, other attributes of an Open Data modality can have significant impacts on adoption and support. Two of these are licensing/IPR and data preservation and management. These topics are addressed in the following sections.

5.2 Licensing Options and Policy

5.2.1 The US Bayh-Dole Act and Intellectual Property

The Bayh-Dole Act or Patent and Trademark Law Amendments Act, passed in 1980, is US legislation that deals with intellectual property arising from government-funded research. This was a particularly important piece of legislation for universities and other not-for-profit organizations receiving funding from the federal government in that the act provided these entities with control over the intellectual property arising from such federal funding (U.S. Congress 1980). The federal government retained a non-exclusive, non-transferable, irrevocable, paid-up license to practice or have practiced on its behalf throughout the world. Through this legislation, the university and the inventor owned the intellectual property rights. The oft-argued idea that since the research was sponsored by the federal government, the rights belonged to all, is no longer a valid, legal point of view in the United States. This was a revolutionary idea and has had a profound impact upon university access to the intellectual property created through federal funding; licensing now brings significant annual returns to many US research universities. The entity supported by funding from the federal government holds the intellectual property rights for work done by that entity. Subcontracts, for example, may transfer the potential for ownership down the chain to where the work has been conducted. The Act has provided US universities the freedom to manage intellectual property directly and various approaches, including licensing, for opening access to data have followed. There are also subsidiary agreements which impact ownership of IPR. Many times, universities will require their staff and employees to sign IPR agreements giving the University exclusive right to the IPR rather than sharing it with the inventor. Variations in these relations introduce variability when looking to develop uniform practices for open data implementation.

5.2.2 The BSD License

While the data from an observatory or an investigator may be open and available, it's important to consider formal approaches to protect both the user and provider through the use of licenses. An early approach was the Berkeley Standard Distribution (BSD) of the Unix operating system. Attributions to the distribution are still quite important; for example, the BSD license is a major portion of the Apple OS X Operating System. While BSD was originally intended to license open software, the NSF Ocean Observatories Initiative (OOI) Cyberinfrastructure may use BSD to license Open Data as well. The permissive license places minimal restrictions on how the data/software can be used and how it is redistributed. For our purposes, the modern 2-clause license ('Simplified BSD' license or 'FreeBSD' License) is most instructive and is consistent with the GNU General Public License (GPL) (FSF 2014). In this license the first clause contains the word 'Copyright,' the copyright symbol (c) along with a year and name of the organization making the claim. This copyright statement would have to be repeated in redistributions of software or data. The second clause of the 'Simplified BSD' license is a 'hold harmless' clause that indemnifies the copyright holder against any future damage resulting from the use of licensed software or data.

5.2.3 GEOSS

GEOSS has integrated a legacy approach on licensing of Earth observation data and information into a summary white paper (Onoda 2012) for the global observatory community entitled "Legal options for the exchange of data through the GEOSS Data-CORE." The white paper lists four principles:

1. The data are free of restrictions on reuse is required;
2. User registration or login to access or use the data is permitted;
3. Attribution of the data provider is permitted as a condition of use; and
4. Marginal cost recovery charges (i.e., not greater than the cost of reproduction and distribution) are permitted.

Of these four items, the first is the most important and declares that data reuse is unlimited; the remaining three are permitted, but not required. This is similar to the BSD license. However, the GEOSS approach does not include the copyright statement or the hold harmless clause in BSD. In terms of credit to the original data producer and potential liabilities attending the use of the data, the GEOSS statement is wanting.

The GEOSS white paper also correctly notes that copyright or database protection (and software) laws arise automatically; there is no need for copyright to be memorialized by filing or statement. On the other hand, the White Paper notes:

> "Hence, either express legislative or regulatory action, or a waiver of all rights through a private law alternative is needed to make the reuse and redissemination of data unrestricted."

The BSD license is one approach to removing the constraints of copyright although, as shown above, an organization continues to hold the copyright, but provides conditions for use of the data or software.

Much of the discussion up to this point deals with the United States in which intellectual property, copyright and patents are federal government functions. This is not the case in Europe where it is important not only to comply with European Union law, but with local state law as well. Creative Commons (2014) provides a means for dealing with the multiplicity of (EU and US) laws.

5.2.4 Creative Commons

Creative Commons was included in the GEOSS Summary White Paper (Onoda 2012) and details are available at the Creative Commons web site (Creative Commons 2014). Palfrey and Gasser (2012) support the idea of using this approach to manage intellectual property by taking a permissive approach for exchanging works across systems, applications and components. There are six Creative Commons licenses available (Creative Commons 2014), extending from a rights management system much like BSD license, to much more constrained ones, which prevent modification of the software or data or its use for commercial purposes. Unlike the BSD license that was initially developed specifically for software, the creative commons license suite was designed from the start to apply to a variety of works that can be covered by copyrights.

The Creative Commons (Creative Commons 2014) public copyright licenses incorporate a unique and innovative “three layer” design. Each license begins as a traditional legal tool, that is the Legal Code layer of each license. But since most creators, educators, and scientists are not lawyers, the licenses are also available in a plain language format known as the ‘Commons Deed’. The Commons Deed is a handy reference for licensors and licensees, summarizing and expressing some of the most important terms and conditions. The Deed itself is not a license, and its contents are not part of the Legal Code itself.

The final layer of the license design recognizes that software, from search engines to office productivity to music editing, plays an enormous role in the creation, copying, discovery, and distribution of works. In order to make it easy for the Web to know when a work is available under a Creative Commons license, a “machine readable” version of the license is provided in a summary of the key freedoms and obligations written into a format that software systems, search engines, and other kinds of technology can understand. Thus, there is a standardized way to describe licenses that software can understand called Creative Commons Rights Expression Language to accomplish this.

6 Data Publication/Data Citation

Going back to the cultural issues initially addressed in section 2, there is limited academic recognition and promotion for collection and publication of data. In the past, there was little incentive for a researcher to make data available. Publication of raw data does not carry the same weight in deciding promotion as papers that include scientific analyses of the data. Promotion criteria do not take into account the innovation and complexity of data acquisition in the ocean’s challenging environments. Many times, data were needed to prepare a peer reviewed research paper, which was essential to further a research career, but there was no benefit to making data supporting the publication accessible. The effort to produce data was not highly rated, yet data are the basis of progress in science and research. Sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and population resources by avoiding duplicate data collection (Piwowar 2011).

Can data publication and data citation offer a solution to some of the human motivation issues discussed in this paper? If so, how would it best be implemented, addressing both questions of “how to publish data under the open access model and how to motivate data collectors and creators?” (Penev 2009) Within the informatics community an interesting question has been raised—should we be using the metaphor “data publication.” It is argued “that there is no widely understood and accepted definition of what exactly Data Publication means.” It was equally clear that “publication” carries many differing implicit assumptions that may not be true (Parsons and Fox 2013). The conclusion was that no one metaphor suits all systems or methods. The term is often used interchangeably with data sharing, but data publishing implies something more. It is a way of using best practices and standards to make sure that data really can be discovered and reused effectively, and that data owners and custodians get the recognition for making datasets public’ (GBIF 2014). In this paper, we define Data Publication as making data freely accessible [or at marginal cost] and permanently available on the Internet along with information as to its trustworthiness, reliability, format and content to enable discovery and re-use. Data Publication can take a number of forms including: Standalone Data Publication; Data Publication by Proxy; Appendix Data; Journal Driven Data Archival; and Overlay Publication (Lawrence et al 2011).

6.1 From Fieldwork to Citation

There are established and/or emerging workflows for selected disciplines that enable the publishing of data and credit via citation mechanisms. However, in many disciplines, researchers are simply not aware of such workflows (WDS 2014) or have the data management support or appropriate training.

In order to motivate research scientists to engage in Open Data models, there must be a clear understanding of the benefits of participation. Such understanding should be based on the end-to-end flow of information from fieldwork to citation. As shown in Figure 1, key elements of the flow are identified and should be primary areas for collaboration and improvement in the emerging transition to Open Data.

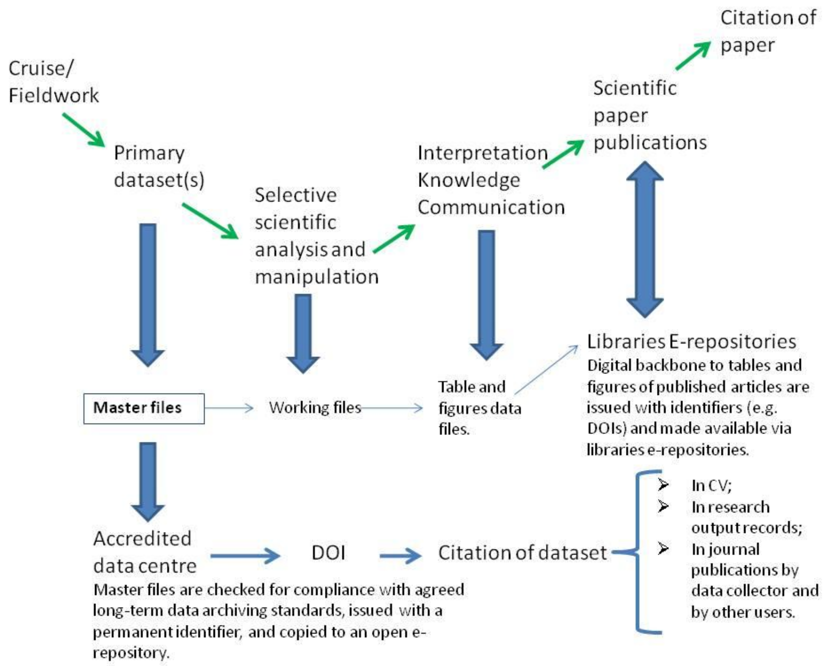
Citation of paper
Cruise/ Fieldwork
Scientific paper publications
Primary dataset(s)
Interpretation Knowledge Communication
Selective scientific analysis and manipulation
Libraries E-repositories
Digital backbone to tables and figures of published articles are issued with identifiers (e.g. DOIs) and made available via libraries e-repositories.
Master files
Working files
Table and figures data files.
In CV;
In research output records;
In journal publications by data collector and by other users.
Accredited data centre
DOI
Citation of dataset
Master files are checked for compliance with agreed long-term data archiving standards, issued with a permanent identifier, and copied to an open e-repository.

Figure 1. End-to-end flow of data and information going from collection to publishing of data. SCOR/MBLWHOI/IODE (2014).

There has been much discussion in the community on incentives for researchers to publish their data: data repositories, citation increase, DOIs, Funding Agency mandate compliance, kudos and recognition for the data creator et cetera (Costello 2009, 2013; Piwowar, 2011, 2013; Sayogo and Pardo 2012; Sears 2011; Tenopir et al 2011). Yet, "despite policies and calls for scientists to make data available, this is not happening for most environmental- and biodiversity-related data because scientists' concerns about these efforts have not been answered and initiatives to motivate scientists to comply have been inadequate" (Costello 2009). Survey results from Cragin et al. (2010) indicate the many and varied concerns of researchers including their perceptions of private sharing versus public sharing and real issues with misuse of data. Positive responses to these concerns will go some way to providing incentives to facilitate open data.

The discussion reflects that data collected by scientists and data managers, whether generated from research or operational observations, are not always deposited in national or international data repositories/archives or deposited in a format that makes them retrievable and reusable. Scientists rarely have the skills or resources needed to prepare all their data for public sharing (RIN 2008). Even when submitted, the data often lack a bare minimum of metadata. The problem is in part cultural.

As research careers are heavily dependent upon journal publications and related citations, researchers wish to hold on to "their" data as long as possible to generate more research papers. In addition, the portability of computing power (researchers can easily store years of data on their laptop) and researchers frequent lack of the most basic data management and preservation practices, makes data unavailable and constantly at risk of being lost. Added to this are the restrictions imposed by the institution or government concerning sensitive data that reduce the "open and free" exchange and access to data (see section on Data Access Models). Below we highlight early project work for data publication: within Institutional/Thematic Repositories and Data Centres, specifically implemented to assist data publication within the ocean science community.

Data publication by deposit into institutional data repositories (IRs) may ensure provenance, permanence, attribution and metadata; at present IRs do not guarantee the scientific quality of published data which requires domain experts more likely found in domain focused data centres. Organizations such as the MBLWHOI Library (MBLWHOI Library 2014) serving the Woods Hole scientific community and supporting the Biological and Chemical Oceanography Data Management Office (BCO-DMO) (see Use Case 2); Lamont's Integrated Earth Data Applications (IEDA 2014) and Scripps' Geological Data Center for geology which includes oceanography (GDC 2014), are among 'early players' participating in World Data System/Research Data Alliance groups (WDS 2014) that are developing standards to be used across all disciplines. Repositories and services established specifically for data publication/sharing and preservation are under development including: Living Atlas of the World (Environmental Systems Research Institute 2014); Planet OS (Planet OS 2014; UKDS Re-Share 2014). The Ocean Observatories Initiative (OOI 2014) confirms it intends to maintain access to data collected over the 25-30 year life of the observatory.

In the research community, peer review is the accepted process for evaluating the quality of scientific work. Acceptance of a community-agreed peer review procedure for data publication approaching those expected for paper publications is not currently available, yet it is essential, to support reliable and trustworthy data publication and to offer data creators the kudos for promotion. Emerging open access data journals that publish papers on the management of data and articles on original research data (sets) are now offering peer review including data e.g. Biodiversity Data Journal and other Pensoft journals (BDJ 2014), Data Science Journal from CODATA (CODATA 2009), Earth System Science Data (Copernicus Publications 2014), F1000Research (2014), Geoscience Data Journal (Wiley 2014), Scientific Data (Nature Publishing Group 2014) and the new AGU Earth and Space Science journal (ESS 2014).

A need for a comprehensive peer review of data publication was stated in Parsons et al, (2010) and the first steps for a formal data peer review were given in 2011 by Lawrence et al. (Dusterhus 2014). A new statistical scheme for quality evaluation by domain experts is also described by Dusterhus including discussion not only on the quality of data but also the quality of the metadata to provide optimal description for discovery and reuse and interestingly, the quality and availability of the reviewers. Blog comments from the 9th International Digital Curation Conference, Feb 2014 Breakout Session, evidences that the discussion on data validation (and peer review) still abounds with ideas and is on-going (Kratz, 2014). Data peer review for Ocean Science is in the same place as other disciplines – there is 'processing' and quality control at the data centres, but not traditional external peer review.

The advent of Funding Agency mandates for Open Data, such as the requirement in the National Science Foundation Data Management Plan (NSF 2010) and the European Commission's recent recommendation for open access to scientific publications and data within Horizon 2020 (EU 2013) and Research Councils UK (2014) is expected to stimulate authors to make data available. Data repositories and data journals providing citation metrics will offer evidence of compliance and multiple venues for data publication.. In addition to standard search engines, secondary services like the Thomson Reuters Data Citation Index (Thomson Reuters 2014), will facilitate discovery, use and attribution of datasets and data studies by connecting researchers and data repositories around the world.

A successful early example of motivating data publication is the cooperative work of four organizations: The Marine Biological Laboratory/Woods Hole Oceanographic Institution (MBLWHOI) Library; the Scientific Committee on Oceanic Research (SCOR); the British Oceanographic Data Centre (BODC); and the International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission (IOC). These organizations have developed and executed a pilot project (SCOR-MBLWHOI-IODE 2014) related to two Use Cases:

1. Data held by data centres are packaged and served in formats that can be cited.
2. Data related to traditional journal articles are assigned persistent identifiers referred to in the articles and stored in institutional repositories;

The goal of the Use Cases has been to identify best practices such as Open Access Initiative (OAI) standards for web content; metadata—Dublin Core, Darwin Core; vocabularies and the ability to add other standards for tracking data provenance and clearly attributing credit to data creators/providers so that researchers will make their data accessible. The assignment of persistent identifiers, specifically Digital Object Identifiers (DOI 2014), enables accurate data citation. The project is also investigating Uniform Resource Identifiers (W3C 2001) and NameIDs. The two project data repositories are meant to be complementary to national and international (e.g., IODE, NODCs, ICSU World Data System and thematic data centres, rather than a replacement. A “cookbook” has been published (Leadbetter et al 2013) that provides extensive instructions and guidelines to scientists as well as the data publication process to repository managers. It identifies that some form of infrastructure and process must be created to motivate and support data publication.

6.2 Use Cases

Two uses cases were developed as exemplars in ocean science for the discussion of data publication and review. The first is the BODC Published Data Library (PDL) (BODC 2014) and the second is the work of MBLWHOI Library Woods Hole Open Access Server (WHOAS). Both WHOAS and PDL are indexed by Thomson Reuters Data Citation index, enabling researchers to gain metrics for their data publication. For the purposes of this paper only the MBLWHOI project is described in detail. Other similar repository models for data publication include Dryad (Dryad 2014) and Pangaea (Pangaea 2014). The sharing of repository records through harvesting also provides greater exposure for data exchange.

6.2.1 Use Case 1

The Published Data Library is implemented by the British Oceanographic Data Centre (BODC 2014). It provides snapshots of specially chosen datasets that are archived using rigorous version management. The publication process exposes a fixed copy of an object and then manages that copy in such a way that it may be referred to over an indefinite period of time. Using metadata standards adopted across NERCs Environmental Data Centres (NERC 2014), the repository assigns DOIs, obtained from the British Library/DataCite, to appropriate datasets.

6.2.2 Use Case 2

The MBLWHOI Library has successfully assigned DOIs to a number of datasets associated with published articles. In the ideal scenario, the DOI(s) should be assigned to the dataset(s) before the article is published, but within the framework of the project there is the ability to retroactively link data to articles after publication. The system has been in operation for over three years, and there is growing interest in the work. Author reaction has been very positive. “This was much easier than trying to deposit data with a publisher”; “The data will be in an open access environment, not owned by publishers”; “Great to know that if my data on my hard disks gets lost at least I have the library copy”. It is interesting to note the bias against publishers, something that should be addressed as a broad-spectrum solution for sustainability evolves.

Scientists are now becoming aware that DOIs offer the means to easily cite their datasets and gain important citation metrics. Librarians have been using DOIs for years and they are now becoming the de facto standard for data citation within data repositories and institutional repositories (commonly universities) and are being facilitated in such services as NASA's EOSDIS (NASA 2014), Pangaea, Dryad, et cetera. Many current data projects register their DOIs with DataCite (DataCite 2014), an organization that is working to develop standards to foster data access and reuse. The MBLWHOI Library registers DOIs with CrossRef (PILA 2013). The Library began assigning DOIs before DataCite existed and many of the major publishers use CrossRef, but there are a number of DOI Registry Agents.

Publishers are now acknowledging the importance of datasets supporting and within published articles. Nature Publishing developed a platform in 2012 and in 2013 PLoS announced a new data sharing policy (Silva 2014). Supporting data made available in a data repository provides publishers with a safe and easy means of linking the dataset to the published article without them having to publish an annex, deal with data on DVDs, or setting up their own data repository.

Many publishers have identified a specific repository for this purpose (in the medical sciences, publishers use PubMed and in fact are required to do so by such Funding Agencies as the National Institutes of Health and the Wellcome Trust. In ocean science, funding agencies like the UK Natural Environment Agency (NERC) require all data created through their grants to be deposited in the British Oceanographic Data Centre but at present there is no one repository designated by publishers for ocean data. Many publishers do not yet have an identifiable policy dealing with supporting datasets (JoRD 2013), though this is now changing with publishers forging new partnerships to store supplemental data; for example, Taylor and Francis Journals (and others) are now using figshare (figshare 2014) who will host the supplemental data as well as provide a widget that will enable Taylor and Francis users to view data in the articles in the browser alongside the content (Research Information 2014).

Because of the assignment of DOIs, Elsevier Publishing sought collaboration with the MBLWHOI Library. Article records in ScienceDirect (ScienceDirect 2014) now contain links to datasets deposited in the Woods Hole Open Access Server (WHOAS 2014) that are associated with Elsevier articles. This system works for DOIs assigned before and after article publication and a WHOAS statement covers copyright, "All Items in WHOAS are protected by original copyright, with all rights reserved, unless otherwise indicated." In addition some depositors request a specific Creative Commons License (Creative Commons 2014). The WHOAS system of linking data to the articles in ScienceDirect was implemented in May 2012.

Another outcome of the project includes tools and procedures developed by the MBLWHOI Library and the NSF funded Biological and Chemical Oceanography Data Management Office (BCO-DMO 2014) to automate the ingestion of data and related metadata from BCO-DMO into the WHOAS Institutional Repository (IR). WHOAS is built on the DSpace platform (DuraSpace 2014). The system also incorporates functionality for BCO-DMO to request a Digital Object Identifier (DOI) from the Library. This partnership allows the Library to work with a trusted data repository to ensure high quality data while the data repository uses library services and is assured that a permanent archived copy of the data is associated with the persistent DOI. Feedback from BCO-DMO is very positive. The Data Manager reports that the most sought after functionality is the DOI and the ability to cite the data. This use case has demonstrated that data can be successfully deposited into a library institutional repository and that the assignment of DOIs is an effective way to enable data citation.

The Library is also participating in an NSF Grant that will result in WHOAS content being published as Linked Open Data which will expose relationships between DSpace repository content and other data sources. Linked Open Data enables knowledge discovery, sharing and integration. Exposing linked data is a concept continuing to emerge. Tim Berners-Lee's vision in 2009 (Berners-Lee 2009) was to "build a web for open, linked data that could do for numbers what the Web did for words . . . unlock our data and reframe the way we use it together."

6.3 Data Citation

Previously, researchers have not really understood how to cite data (or compile meaningful metadata) and "full citation of data is not currently a normative behaviour in scholarly writing" (Mooney and Newton 2012). However, the introduction of DOIs for data sets has been a positive encouragement welcomed by the research and informatics community. The advent of a number of Research Data Training online courses, e.g., (MANTRA 2014) are welcome tools for researchers to gain RDM skills.

Citation metrics have been adopted across the sciences as a method to obtain quantitative indicators for the assessment of the quality of research and researchers, as well as the impact of research products. Systems and services such as the Science Citation Index (Thomson Reuters 2013), the h-index (or Hirsch number), or the Impact Factor of scientific journals have been developed to track and record access and citation of scientific publications. These indicators are widely used by investigators, academic departments and administration, funding agencies, and professional societies across all disciplines to assess performance of individuals or organizations within the research landscape and inform and influence the advancement of academic careers and investments of research funding. New data metrics indicating the value and impact of data publications (like those launched by the Data Citation Index in 2012) are needed to raise the value and appreciation of data and data sharing because the missing recognition for data publication in science is seen as a major cause for the reluctance of data producers to share their data (Smit 2010). Calls for data sets to be cited in a conventional manner are widespread and the growing use of persistent DOIs assigned to data sets (e.g., by MBLWHOI, DataCite and Dryad) is a major contribution leading to a call for a central registry resolving various digital identifiers (DOI, URL, URI et cetera) (Costello 2013).

The Research Data Alliance (RDA 2014) supported by the European Commission, the U.S. Government and the Australian Government is likely to have a significant impact on the research data landscape. An overall objective of the ICSU World Data System/Research Data Alliance WDS/RDA Interest Group on Data Publication - Bibliometrics is to "conceptualize data metrics and corresponding services that are suitable to overcome existing barriers and thus likely to initiate a cultural change among scientists, encouraging more and better data citations…" (WDS 2014).

Work among several groups is resulting in recommendations for data citation formats; early examples include Altman and King (2007), the UK Digital Curation Centre (Ball and Duke 2012) and the Federation of Earth Science Information Partners (ESIP 2012). Data Citation Groups have been formed; the Force 11 Data Citation Synthesis Group has released and called for endorsement of the consolidated Joint Declaration of Data Citation Principles, a collaborative effort including such groups as the CODATA-ICSTI Task Group on Data Citation (Force II 2013). Other data citation groups like that within Mendeley (Mendeley 2014) and UK Data Service (UKDS 2014) contribute to the discussion.

"Data publication and data citation is becoming increasingly important to the scientific community, as it will provide a mechanism for those who create data to receive academic credit for their work and will allow the conclusions arising from an analysis to be more readily verifiable, thus promoting transparency in the scientific process" (Lawrence et al 2011). It can create incentives for researchers to make data available with sufficient metadata, to make it discoverable and re-usable, thereby gaining citations. Of course, this is conditional upon institutional management agreeing to use data citation metrics as an element in performance assessment and career advancement decisions. The recent trend of Funding Agencies and Publishers requiring data related to publications to be accessible will accelerate data publication.

7 Recommendations

7.1 Interoperability/Standards Recommendations

Within domains, standards for data formats have developed and are in use so that exchange of data is not greatly impeded. Across domains there is a greater problem where one set of standards and formats is incompatible with another or where there are differing interpretations/implementation of a given standard. For domains, which have only recently come to work with each other, patches are possible and translation programs have been written. The more general solution of having universal standards so that all domains can exchange data is a distant hope, similar in dimension to a universal spoken and written language across the entire world. Yet, spoken communications between people worldwide can be accomplished with at most three to ten languages, English, Mandarin, Arabic, Spanish, Russian, French, German, and Swahili form a short list some of which might serve the universal spoken language requirement. Not everyone is accommodated and most will be communicating with a second or third language, not their native tongue. And this may be as close as we can expect for a universal standard for data formats. Like the adoption of English as a lingua franca, commerce can be a major force in promoting interoperability. The ubiquity of products from a few software ’giants’ has similarly forced compliance with their formats, forming de facto standards enabling a rudimentary form of interoperability. If Microsoft, Adobe, MathWorks, and half a dozen other software firms are considered, standards are developing that permit exchange of data and interoperability, though in many cases awkwardly or inefficiently. Particular solutions, as between two newly interacting communities with ad hoc standards, are the path that is presently recommended (National Research Council 2012; Leadbetter et al, 2013).

Ultimately, the broadly inclusive collaborations across scientific disciplines need a more formal way to make data generally available. Translators for formats must develop as a middleware market. Recent developments in information brokering have been quite encouraging, and demonstrations with selected user scenarios and communities have pointed to significant benefits (Nativi et al. 2013). Further development, implementation and uptake of brokering middleware is recommended as an important step forward. The ocean science community, with its wide and multi-disciplinary diversity is an excellent test bed for such implementation demonstrations.

The ability of users to feel comfortable in a cross-domain environment is essential to further collaboration and addressing the complexities of global issues. Thus, outreach and capacity building are needed to aid users in accessing data and the appropriate support services. Such activities should be built into the adoption and acceptance of Open Data.

7.2 Governance and Business Model Recommendations

The costs for maintaining the research infrastructure, data management and publishing require significant investments. Even relatively small elements of the system such as the peer review and publishing process, using volunteer reviewers, still requires substantial financial resources. Government support is pervasive throughout the research environment, covering infrastructure, salaries, university research activities, data management, publishing and community exchanges. Much of the support is built upon rights or business frameworks that have adapted to the pre-Open Data model. This covers many things such as IPR for universities and subscription-based support for publication. For example, journal support is evolving from user/subscription-based to author-based fees. In this transition to Open Data, the essential attributes of the system of the broad research infrastructure as described in Section 5.1 should be maintained and improved. The social elements such as recognition for work, awarding of grants and career advancement drive uptake of the Open Data paradigm and these motivations should be addressed. This will involve a substantial outreach and education program on advantages of Open Data. It will also mean that impact metrics need to be created, accepted and clearly visible to the community at large.

The fiscal impacts of Open Data must be addressed so that viable business models for key elements of the end-to-end infrastructure can be defined and maintained. By openly using and redistributing data, some of the assumptions underlying the current operating practices will need to be adapted. Clearly defining the boundary conditions for the Open Data environment will speed the process. Simply stating that ‘all data’ will be ‘open’, without widespread, consistent adoption and without adjusting the balance of the system will undercut viability of the Open Data Policy.

In its implementation, Open Data must improve the efficiency and impacts of scientific research. This will be achieved when Open Data implementation and Policy:

- Ensures sustainability;
- Preserves the peer review attributes of science and of publications;
- Assures scientists of recognition for their scientific research;

- Maintains data attributes such as provenance, metadata, quality attributes, etc.;
- Allows easy discovery and access to data and information, particularly supporting cross discipline research;
- Supports IPR and licensing protocols;
- Consistent with national and international policies;
- Motivates participation and contributions;
- Minimizes impacts on existing discipline-specific systems;
- Works across physical, social and economic sciences; and
- Promotes Access and use by the public and policy makers.

Within organizations that support research such as NSF, NIH, etc., metrics should be established to monitor progress in these areas. Furthermore, each organization should establish a process with broad stakeholder representation to make recommendations on issues, both for implementation and operations. Policies should be adopted that support the sustainability of Open Data over the long term.

7.3 Data Publication/Data Citation Recommendations

1. Data Publication that enables data citation can certainly be an incentive to make data more accessible. The associated functionality to deposit data safely and securely should be attractive to the researcher and of course the additional citation of the data associated with a research paper will add value to these data as an essential component of research output. In addition, data publication and data citation can create incentives for researchers, provided that institutional management use the data citation metrics as an element in performance assessment and career advancement decisions.
2. An accepted peer review methodology for datasets and/or data repositories has been discussed in the data management community at meetings such as the 2012 Fall American Geophysical Union Meeting. This is an essential step. Discussions (National Research Council (2012) and Harley et al (2010)) should also consider implementation of solutions to issues of time, credit, and peer review as compared between 12 disciplines: Anthropology, Biostatistics, Chemical Engineering, Law and Economics, English-language Literature, Astrophysics, Archaeology, Biology, Economics, History, Music, and Political Science.
3. A call for all journal publishers to have a clearly stated data policy regarding supplemental material and related datasets would eliminate confusion for authors and hopefully lead to the establishment of standards across publishers.
4. Research Data Management training should be included in University curricula.
5. A consistent, predictable policy on publishing costs and access costs should be addressed for the Open Data environment assuring that the peer review system and publication quality will be maintained.
6. Adoption of Digital Object Identifiers or equivalent “globally unique persistent identifiers” should be expanded and widely implemented. This includes DOIs as part of an important and sustainable infrastructures registering and distributing data sets. This requires a long-term commitment to ensuring that data are viable.

Collaboration between international repositories of ocean science and other data should be encouraged both to improve efficiency and reduce costs. A working group under the NSF OceanObsNetwork RCN exists to support such collaboration.

Acknowledgements

The authors would like to thank other members of the NSF Research Coordination Network “OceanObsNetwork”: Milton Kampel; Takeshi Kawano; Fred Maltz; Michael McCann; Benoit Pirenne; Peter Pissierssens; Iain Shepherd; Christoph Waldmann; and Albert Williams III, who contributed to the report that this paper summarizes and enhances (Pearlman et al 2013). The authors acknowledge the support of the National Science Foundation through Grant Award No. OCE-1143683.

References (Revised with additions)

Allcock W, Bresnahan J, Kettimuthu R, Link M, Dumitrescu C, Raicu J, Foster I (2005) The Globus striped GridFTP framework and server. In Proceedings of the 2005 ACM/IEEE conference on Supercomputing p 54 IEEE Computer Society, 2005.

Allinson J (2006) OAIS as a reference model for repositories: an evaluation. Report, UKOLN, University of Bath. http://eprints.whiterose.ac.uk/3464/. Accessed September 19, 2014.

Altman M, King G (2007) A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). http://www.dlib.org/dlib/march07/altman/03altman.html. Accessed 19 September 2014.

Australian Government (2009) Government 2.0 Task Force Report. http://www.finance.gov.au/publications/gov20taskforcereport/doc/Government20TaskforceReport.pdf. Accessed 16 September 2014.

Ball A, Duke M (2012) How to cite data sets and link to publications. Edinburgh, UK: Digital Curation Centre. http://www.dcc.ac.uk/webfm_send/525. Accessed 16 September 2014.

BDJ (2014) Biodiversity Data Journal. http://biodiversitydatajournal.com. Accessed 16 September 2014.

BCO-DMO (2014) Biological & Chemical Oceanography Data Management Office. http://www. bco-dmo.org. Accessed 16 September 2014.

Berners-Lee T (2009) The next web. TED 2009 Conference. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html. Accessed 16 September 2014.

Best B, Halpin P, Fujioka E, Read A, Qain S, Hazen L, Schick R (2007) Geospatial web services within a scientific workflow: Predicting marine mammal habitats in a dynamic environment. Ecological Informatics 2(3):210–223. doi: 10.1016/j.ecoinf.2007.07.007.

BODC (2014) Published Data Library. https://www.bodc.ac.uk/data/published_data_library. Accessed 16 September 2014.

Borgman CL (2012) The conundrum of sharing research data. Journal of the Association for Information Science and Technology 63(6):1059–1078, doi:10.1002/asi.22634.

Braunschweig K, Eberius J, Thiele M, Lehner W (2012) The state of open data—limits of current open data platforms. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S (eds) Proceedings of the 21st World Wide Web Conference 2012, (WWW 2012), Lyon, France.

Busse S, Kutsche RD, Leser U, Weber H (1999) Federated information systems: concepts, terminology and architectures. Tech. Rep., Technical University Berlin.

Carpenter S et al. (2009) Accelerate synthesis in ecology and environmental sciences. BioScience 59(8):699–701 doi: 10.1525/bio.2009.59.8.11

Cocco M (2012) Research infrastructure and e-science for data and observatories on earthquakes, volcanoes, surface dynamics and tectonics. ICRI2012, International Conference on Research Infrastructures. http://ec.europa.eu/research/infrastructures/pdf/workshop_october_2011/16_esfri_epos_cocco.pdf. Accessed 16 September 2014.

CODATA (2009) Data Science Journal. http://www.codata.org/dsj. Accessed 16 September 2014.

Copernicus (2014) Copernicus: The European Earth Observation Programme. http://www.copernicus.eu. Accessed 16 September 2014.

Copernicus Publications (2014) Earth System Science Data: the data publishing journal. http://earth-system-science-data.net. Accessed 16 September 2014.

Costello M (2009) Motivating online publication of data. BioScience 59(5):418–427. doi: 10.1525/bio.2009.59.5.9.

Costello M, Bouchet P, Boxshall G, Fauchald K, Gordon D, et al. (2013) Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. PloS one 8(1): e51629. doi: 10.1371/journal.pone.0051629.

Costello M, Michelner WK, Gahegan M, Zhang Z-Q, Bourne PE (2013) Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution 28(8):454–461.

Costello M, Appeltrans W, Bailly N, Berendsohn W, Jong Y, Edwards M, Froese R, Huettmann F, Los W, Mess J, Segers H, Bisby F (2014) Strategies for the sustainability of online open-access biodiversity databases. Biological Conversation 173:155–165.

Costello M, Wieczorek J (2013) Biological Conversation 173:68–73 doi: 10.1016/j.biocon.2013.10.018.

Cragin MH , Palmer CL, Carlson JR., Witt M (2010) Data sharing, small science and institutional repositories. Philosophical Transations of the Royal Society A 368:4023–4038.

Creative Commons (2014) Creative commons license. http://creativecommons.org/licenses. Accessed 16 September 2014.

CUAHSI (2013) CUAHSI Water Data Center. http://wdc.cuahsi.org. Accessed 16 September 2014.

Datacite (2014) Datacite. https://www.datacite.org. Accessed 16 September 2014.

DataNet (2014) DataNet Federation Consortium—collaboration environments for data drivenscience. http://datafed.org. Accessed 16 September 2014.

DataONE (2014) NSF data observation network for earth (DataONE). https://www.dataone.org/about. Accessed 16 September 2014.

DOI (2014) Digital object identifier. http://www.doi.org. Accessed 16 September 2014.

Dryad (2014) Dryad Digital Repository. http://datadryad.org. Accessed 16 September 2014.

DuraSpace (2014) DSpace. http://www.dspace.org. Accessed 16 September 2014.

Dusterhus A, Hense A (2014) Automated quality evaluation for a more effective data peer review. Data Science Journal 13:67–78.

Earth Observations (2013) GEO BON—biodiversity observation network http://www.earthobservations.org/geobon.shtml. Accessed 16 September 2014.

Environmental Systems Research Institute (2014) Living Atlas of the World. http://doc.arcgis.com/en/living-atlas. Accessed 21 September 2014.

ESIP (2012) Federation of Earth Science Information Partners. http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations. Accessed 16 September 2014.

ESS (2014) Earth and Space Science. http://agupubs.onlinelibrary.wiley.com/agu/journal/10.1002/%28ISSN%292333-5084/. Accessed 17 September 2014

EU (2006) Communication from the Commission to the Council and the European Parliament: Interoperability for pan-European government services. http://www.epsos.eu/uploads/tx_epsosfileshare Communication-on-Interoperability_01.pdf. Accessed 16 September 2014.

EU (2013) Guidelines on open access to scientific publications and research data in Horizon 2020. Version one. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Accessed 16 September 2014.

European Commission (2011) Digital agenda: turning government data into gold. Press release. http://europa.eu/rapid/press-release_IP-11-1524_en.htm. Accessed 16 September 2014.

F1000Research (2014). http://f1000research.com. Accessed 16 September 2014.

FGDC (2014) National spatial data infrastructure (NSDI).http://www.fgdc.gov/nsdi/nsdi.html. Accessed 16 September 2014.

figshare (2014) figshare. http://figshare.com. Accessed 16 September 2014.

Folkman M, Liao L, Jarecke P (2001) EO-1/Hyperion hyperspectral imager design, development, characterization, and calibration, Proc. SPIE 4151, Hyperspectral Remote Sensing of the Land and Atmosphere, 40 (February 8, 2001); doi: 10.1117/12.417022;

Force II (2013) Force II: The future of research communication and scholarship, Joint Declaration of Data Citation Principles. http://www.force11.org/datacitation. Accessed 16 September 2014.

FSF (2014) Free software foundation. http://www.fsf.org. Accessed 16 September 2014.

Gallagher J, Potter N, Sgouros T, Hankin S, Flierl G (2007) The data access protocol—DAP 2.0. NASA ESE-RFC-004.1.1. https://earthdata.nasa.gov/our-community/esdswg/standards-process-spg/rfc/esds-rfc-004-dap-20. Accessed 16 September 2014.

GBIF (2014) Global Biodiversity Information Facility. http://www.gbif.org. Accessed 16 September 2014.

GDC (2014) Geological Data Center, Scripps Institution of Oceanography. http://gdc.ucsd.edu/. Accessed 17 September 2014

GEO/CEOS (2008) GEO/CEOS workshop on quality assurance of calibration & validation processes: Establishing an operational framework. http://qa4eo.org/workshop_washington08.html. Accessed 16 September 2014.

GeoViQua (2007) GeoViQua: QUAlity aware VIsualization for the global earth observation system of systems. http://www.geoviqua.org. Accessed 16 September 2014.

Grassle, J.F. 2000. The Ocean Biogeographic Information System (OBIS): An on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context.Oceanography 13(3):5–7 doi: 10.5670/oceanog.2000.01.

Guess A (2013) Japan embraces open data, launches multiple open projects. http://semanticweb.com/japan-embraces-open-data-launches-multiple-open-projects_b35158. Accessed 16 September 2014.

Hankin S, Blower J, Carval Th, Casey K, Donlon C, Lauret O, Loubrieu T, Srinivasan A, Trinanes J, Godoy O, Mendelssohn R, Signell R, De La Beaujardiere J, Cornillon P, Blanc F, Rew R, Harlan J (2010) NETCDF-CF-OPENDAP : standards for ocean data interoperability and object lessons for community data standards processes, Oceanobs 2009, Venice Convention Centre, 21-25 septembre 2009, Venise, publication date 2010-12-23, http://archimer.ifremer.fr/doc/00027/13832/10969.pdf. Accessed 16 September 2014.

Harley D, Kryzys Acord S, Earl-Novell S, Lawrence, S, Judson King C (2010) Assessing the future landscape of scholarly communication: An exploration of faculty values and needs in seven disciplines. Center for Studies in Higher Education, UC Berkeley. http://escholarship.org/uc/cshe_fsc. Accessed 17 September 2014.

IEDA (2014) Integrated Earth Data Applications. http://www.iedadata.org. Accessed 16 September 2014.

INSPIRE (2014) Infrastructure for spatial information in the European Community (INSPIRE). http://inspire.ec.europa.eu. Accessed 16 September 2014.

IODE (2014) International Oceanographic Data and Information Exchange. http://www.iode.org,. Accessed 16 September 2014.

JoRD (2013) JoRD: Journal research data policy bank project. http://jordproject.wordpress.com. Accessed 16 September 2014.

Klump J, Bertelmann R, Brase J, Diepenbroek M, Gross, H, Hock H, Lautenschlager M, Schingler W, Sens I, Wachter J (2006) Data publication in the Open Access Initiative, Data Science Journal 5:79–83.

Kozak M, Hartley J (2013) Publication fees for open access journals: Different disciplines—different methods. Journal of the American Society for Information Science and Technology 64 (12). doi:10.1002/asi.22972.

Kratz J (2014) Fifteen ideas about data validation (and peer review). Data Pub [Blog], http://datapub.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review. Accessed 16 September 2014.

Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C, et al. (2011) The Development of Open Access Journal Publishing from 1993 to 2009. PloS one 6(6):e20961 doi:10.1371/journal.pone.0020961.

Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S (2011) Citation and peer review of data: moving towards formal data publication. International Journal of Digital Curation 6(12):4–37.

Lavoie B, (2008) The Open Archival Information System Reference Model: Introductory Guide. Microform & Imaging Review. 33(2):68–81, doi: 10.1515/MFIR.2004.68.

Leadbetter A, Raymond L, Chandler C, Pikula L, Pissierssens P, Urban E (2013) Ocean Data Publication Cookbook. Paris: UNESCO, 39pp. (Intergovernmental Oceanographic Commission Manuals and Guides 64). http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10574. Accessed 2014-03-09.

Lecomte P, Stensaas G (2009) Overview of progress towards a data quality assurance strategy to facilitate interoperability. http://www.earthobservations.org/documents/committees/adc/200909_11thADC/ DA-09-01a%20QA4EO.pdf. Accessed 22 April 2014.

MANTRA (2014) Research Data MANTRA.http://datalib.edina.ac.uk/mantra. Accessed 15 September 2014.

Marshall P, Tufo H, Keahey K, La Bissoniere D (2012) Architecting a Large-scale Elastic Environment-Recontextualization and Adaptive Cloud Services for Scientific Computing. In Proceedings of ICSOFT:409–418.

MBL WHOI Library (2014) http://www.mblwhoilibrary.org. Accessed 15 September 2014.

Mendeley (2014) Mendeley. http://www.mendeley.com Accessed 15 September 2014.

Mooney H, Newton M.P. (2012) The anatomy of a data citation: discovery, reuse and credit. Journal of Librarianship and Scholarly Communication, 1(1):eP1035.

NASA (2014) EOSDIS: NASA's earth observing system data and information system. https://earthdata.nasa.gov. Accessed 15 September 2014.

National Research Council (2012) For attribution – Developing data attribution and citation practices and standards. National Academies Press, Washington, DC.

Nativi S, Craglia M, Pearlman J (2012) The brokering approach for multidisciplinary interoperability: A position paper. International Journal of Spatial Data Infrastructure 7:1–15.

Nativi S, Craglia M, Pearlman J (2013) Earth science infrastructures interoperability: The brokering approach. Journal of Selected Topics in Applied Earth Observation and Remote Sensing 6:1118–1129, doi: 10.1109/JSTARS.2013.2243113.

Nature Publishing Group (2014) Scientific Data. http://www.nature.com/sdata. Accessed 15 September 2014.

Neilsen M (2011) Reinventing Discovery: The New Era of Networked Science. Princeton University Press.

NERC (2014) Data centres. http://www.nerc.ac.uk/research/sites/data. Accessed 15 September 2014.

NOAA IOOS (2014) Quality assurance of real time ocean data, QARTOD. http://www.ioos.noaa.gov/qartod/welcome.html. Accessed 15 September 2014.

NSB (2011) NSB 11-79 digital research data sharing and management: report of the Task Force on Data Policies. Tech. rep. National Science Board.

NSF (2010) National science foundation data management plan. http://www.nsf.gov/bfa/dias/policy/dmp. jsp. Accessed 15 September 2014.

NSF (2014) National Science Foundation Directorate for Geosciences: Earth Cube. http://www.nsf.gov/geo/earthcube/. Accessed 22 April 2014.

OGC (2014) Open Geospatial Consortium. http://www.opengeospatial.org. Accessed 15 September 2014.

OneGeology (2014) OneGeology.http://www.onegeology.org, Accessed 24 April 2014.

Onoda M (2012) GEOSS Data sharing principles and action plan. Workshop on GMES Data and Information Policy, Brussels http://ec.europa.eu/enterprise/newsroom/cf/_getdocument.cfm?doc_id=7140. Accessed 15 September 2014.

OOI (2014) Ocean Observatories Initiative. http://oceanobservatories.org/. Accessed 18 September 2014

Open Knowledge Foundation (2012) The open data handbook. http://opendatahandbook.org/en/what-is-open-data. Accessed 15 September 2014.

Palfrey J, Gasser U (2012) Interop: The Promise and Perils of Highly Interconnected Systems. Basic Books.

Pangaea (2014) Pangaea: Data publisher for the earth & environmental sciences. http://www.pangaea.de. Accessed 15 September 2014.

Parsons MA, Fox P (2013) Is data publication the right metaphor? Data Science Journal 12:WDS32–WDS46.

Parsons MA, Duerr R, Minster J-B (2010) Data citation and peer review. Eos: Transactions American Geophysical Union 91(34):297–299.

Pearlman J, Shibasaki R (2008) Guest editorial: Global earth observation system of systems. IEEE Systems Journal 2(3):302–303. doi: 10.1109/JSYST.2008.928859.

Pearlman J, Williams A, Simpson P (eds) (2013) Report of the Research Coordination Network: RCN OceanObs Network: Facilitating open exchange of data and information. NSF/Ocean Research Coordination Network Tech Rep. 46 pp.

Penev L, Erwin T, Mille, J, Chaqvan V, Motitz T, Griswold C (2009) Publication and dissemination of dataset in taxonomy: ZooKeys working example. ZooKeys 11:1-8.

PILA, Inc (2013) CrossRef. http://www.crossref.org. Accessed 15 September 2014.

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PloS one 2(3):e308 doi:10.1371/journal.pone.0000308.

Piwowar H (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. PloS one 6(7):e18657.

Piwowar HA and Vision TJ (2013) Data reuse and the open data citation advantage. PeerJ 1:e175.

Planet OS (2014) Planet OS: Big Data Platform for Multi-Sensor and Machine Data. https://planetos.com/. Accessed 21 September 2014.

President Barack Obama (2013) Memorandum on Open Data Policy–Managing Information as an Asset (May 9, 2013). http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed 17 September 2014.

Reichman O, Jones M, Schildhauer M (2011) Challenges and Opportunities of Open Data in Ecology. Science 331.6018:703–705 doi: 10.1126/science.1197962.

Research Data Alliance (2014). Research data sharing without barriers.https://www.rd-alliance.org. Accessed 15 September 2014.

Research Councils UK (2014) http://www.rcuk.ac.uk/research/datapolicy. Accessed 15 September 2014.

Research Information (2014) Taylor & Francis partners with figshare for supplementary data.http://www.researchinformation.info/news/news_story.php?news_id=1485. Accessed 15 September 2014.

Research Information Network (2008). To share or not to share: publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. http://www.rin.ac.uk/data-publication. Accessed 15 September 2014.

Sayogo DS, Pardo T (2012) Exploring the motive for data publication in open data initiative: Linking intention to action. In: Proceedings of the 45th Hawaii International Conference on System Sciences, IEEE Computer Society.

ScienceDirect (2014) http://www.sciencedirect.com. Accessed 15 September 2014.

SCOR/MBLWHOI/IODE (2014). Data publication/data citation project. http://www.iode.org/index.php?option=com_content&view=article&id=110&Itemid=12. Accessed 15 September 2014.

SeaDataNet (2014) Sea data net. http://www.seadatanet.org. Accessed 15 September 2014.

Sears J (2011) Data sharing effect on article citation rate in paleoceanography. Eos, Trans. AGU, 92, Fall Meet. Suppl., Abstract /IN53B-1628.

Silva L (2014) PLoS new data policy: public access to data. http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2. Accessed 15 September 2014.

Smit E (2010) Preservation, access and re-use of research data. Presented at DataCite Summer Meeting 2010. https://www.datacite.org/datacite_summer_meeting_2010. Accessed 15 September 2014.

SURF (2013) Enhanced publications. Collaborative organisation for ICT in Dutch higher education and research. http://www.surf.nl/en/themes/research/research-data-management/enhanced-publications. Accessed 15 September 2014.

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M (2011) Data sharing by scientists: practices and perceptions. PloS one 6(6):e21101.

Thessen A, Patterson D (2011) Data issues in the life sciences. ZooKeys 150:15–51. doi: 10.3897/zookeys.150.1766.

Thomson Reuters (2014) The Data Citation Index. http://wokinfo.com/products_tools/multidisciplinary/dci. Accessed 11 March 2014.

Thomson Reuters (2013) Science Citation Index. http://science.thomsonreuters.com/cgi-bin/jrnlst/jloptions.cgi?PC=K. Accessed 1 March 2014.

Turnitsa C (2005) Extending the levels of conceptual interoperability model. In: Proceedings IEEE summer computer simulation conference, IEEE CS Press.

US Congress (1980) Bayh-Dole act. Public Law 96–517, also known as the Patent and Trademark Law Amendments Act; enacted by the United States Congress.

UKDS (2014) UK Data Service, Citing Data and Re-Share. http://ukdataservice.ac.uk/media/440282/publishingcitigdata.pdf. Accessed 10 March 2014.

Vision T, (2010) Open Data and the Social Contract of Scientific Publishing. BioScience 60(5):330–1. doi:10.1525/bio.2010.60.5.2.

W3C (2001) URIs, URLs, and URNs: Clarifications and recommendations 1.0. http://www.w3.org/TR/uri-clarification. Accessed 10 March 2014.

WDS (2014) Data Publication Working Group. http://icsu-wds.org/community/working-groups/data-publication. Accessed 10 March 2014.

Whiteside A, Evans JD (2006) Web coverage service implementation specification #06-083r8, version 1.1.0.https://portal.opengeospatial.org/files/?artifact_id=18153, Access 19 May 1024.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PloS one 7(1):e29715.

Whitfield P (2012) Why the provenance of data matters: assessing “fitness for purpose” for environmental data. Canadian Water Resources Journal 37(1):23–36 doi: 10.4296/cwrj3701866.

Whitlock M (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology and Evolution 26(2):61–65. doi: 10.1016/j.tree.2010.11.006.

WHOAS (2014) Woods Hole Open Access Server. https://darchive.mblwhoilibrary.org. Accessed 16 September 2014.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PloS one 7(1):e29715. doi: 10.1371/journal.pone.0029715.

Wiley (2014) Geoscience Data Journal. http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060. Accessed 11 March 2014.

World Meteorological Organization (2014) Information management. http://www.wmo.int/pages/themes/wis/index_en.html. Accessed 11 March 2014.